

The American Economic Review

ARTICLES

Business
Books
GERARD GENNOTTE AND HAYNE LELAND

Market Liquidity, Hedging, and Crashes

LAWRENCE M. AUSUBEL

Insider Trading in a Rational Expectations Economy

JEAN-JACQUES LAFFONT AND JEAN TIROLE

Optimal Bypass and Cream Skimming

DAVID A. BUTZ

Durable-Good Monopoly and Best-Price Provisions

PAUL S. SEGERSTROM, T. C. A. ANANT, AND ELIAS DINOPOULOS

A Schumpeterian Model of the Product Life Cycle

MARC DUDEY

Competition by Choice: The Effect of Consumer Search on Firm Location Decisions

JAMES DEARDEN, BARRY W. ICKES, AND LARRY W. SAMUELSON

To Innovate or Not To Innovate: Incentives and Innovation in Hierarchies

C. Y. CYRUS CHU AND HUI-WEN KOO

Intergenerational Income-Group Mobility and Differential Fertility

MARK M. PITT, MARK R. ROSENZWEIG, AND MD. NAZMUL HASSAN

Productivity, Health, and Inequality in the Intrahousehold Distribution of Food in Low-Income Countries

HEINZ HOLLÄNDER

A Social Exchange Approach to Voluntary Cooperation

MICHAEL DOTSEY The Economic Effects of Production Taxes in a Stochastic Growth Model

MICHAEL C. KEELEY

Deposit Insurance, Risk, and Market Power in Banking

AVNER BAR-ILAN

Overdrafts and the Demand for Money

HENNING BOHN

Tax Smoothing with Financial Instruments

SHORTER PAPERS: A. F. Daughety; D. Levin; P. DeGraba; B. Nahata, K. Ostaszewski, and P. K. Sahoo; M. Schwartz; F. S. Hipple; M. Ogaki; K. A. McCabe, S. J. Rassenti, and V. L. Smith.

DECEMBER 1990

THE AMERICAN ECONOMIC ASSOCIATION

Founded in 1885

Officers

President

GERARD DEBREU

University of California-Berkeley

President-elect

THOMAS C. SCHELLING

University of Maryland

Vice-Presidents

RUDIGER W. DORNBUSCH

Massachusetts Institute of Technology

ALLAN H. MELTZER

Carnegie Mellon University

Secretary-Treasurer

C. ELTON HINSHAW

Vanderbilt University

Editor of The American Economic Review

ORLEY C. ASHENFELTER

Princeton University

Editor of The Journal of Economic Literature

JOHN PENCEL

Stanford University

Editor of The Journal of Economic Perspectives

JOSEPH E. STIGLITZ

Stanford University

Executive Committee

Elected Members of the Executive Committee

GEORGE A. AKERLOF

University of California-Berkeley

ISABEL V. SAWHILL

Urban Institute

STANLEY FISCHER

World Bank

LAWRENCE H. SUMMERS

Harvard University

GREGORY C. CHOW

Princeton University

SUSAN ROSE-ACKERMAN

Yale University

EX OFFICIO Member

ROBERT EISNER

Northwestern University

•Printed at Banta Company, Menasha, Wisconsin.

•Copyright © American Economic Association 1990. All rights reserved.

•No responsibility for the views expressed by authors in this *Review* is assumed by the editors or the publishers, The American Economic Association.

Correspondence relating to advertising, business matters, permissions to quote, subscriptions, and changes of address, should be sent to the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Change of address notice must be received at least six (6) weeks prior to the publication month. A membership or subscription paid twice is automatically extended for an additional year unless otherwise requested.

THE AMERICAN ECONOMIC REVIEW (ISSN 0002-8282), December 1990, Vol. 80, No. 5, is published five times a year (March, May, June, September, December) by the American Economic Association, 2014 Broadway, Suite 305, Nashville, TN 37203. Annual fees for regular membership, of which 30 percent is for a year's subscription to this journal, are: \$44.00, \$52.80, or \$61.60 depending on income. A membership also includes the *Journal of Economic Literature* and the *Journal of Economic Perspectives*. In countries other than the U.S.A., add \$16.00 for extra postage. Second-class postage paid at Nashville, TN and at additional mailing offices. **POSTMASTER:** Send address changes to the *American Economic Review*, 2014 Broadway, Suite 305, Nashville, TN 37203.

THE AMERICAN ECONOMIC REVIEW

Editor

ORLEY ASHENFELTER

Co-Editors

ROBERT H. HAVEMAN
BENNETT T. McCALLUM
PAUL R. MILGROM

Production Editor

J. DAVID BALDWIN

Board of Editors

GEORGE A. AKERLOF
JAMES E. ANDERSON
TIMOTHY F. BRESNAHAN
JOHN Y. CAMPBELL
HENRY S. FARBER
MARJORIE A. FLAVIN
ROBERT P. FLOOD
CLAUDIA D. GOLDIN
JO ANNA GRAY
REUBEN GRONAU
DANIEL S. HAMERMESH
ROBERT J. HODRICK
KEVIN D. HOOVER
KENNETH L. JUDD
JOHN H. KAGEL
JOHN F. KENNAN
DALE T. MORTENSEN
EDGAR O. OLSEN
JOHN G. RILEY
RICHARD ROLL
THOMAS ROMER
DAVID E. M. SAPPINGTON
ROBERT S. SMITH
BARBARA J. SPENCER
RICHARD TRESCH
HAL R. VARIAN
KENNETH WEST
JOHN D. WILSON
LESLIE YOUNG

December 1990

VOLUME 80, NUMBER 5

391
+ 001

Articles

- Market Liquidity, Hedging, and Crashes
Gerard Gennotte and Hayne Leland 999
- Insider Trading in a Rational Expectations Economy
Lawrence M. Ausubel 1022
- Optimal Bypass and Cream Skimming
Jean-Jacques Laffont and Jean Tirole 1042
- Durable-Good Monopoly and Best-Price Provisions
David A. Butz 1062
- A Schumpeterian Model of the Product Life Cycle
Paul S. Segerstrom, T. C. A. Anant, and Elias Dinopoulos 1077
- Competition by Choice: The Effect of Consumer Search on Firm Location Decisions
Marc Dudey 1092
- To Innovate or Not To Innovate: Incentives and Innovation in Hierarchies
James Dearden, Barry W. Ickes, and Larry W. Samuelson 1105
- Intergenerational Income-Group Mobility and Differential Fertility
C. Y. Cyrus Chu and Hui-Wen Koo 1125
- Productivity, Health, and Inequality in the Intra-household Distribution of Food in Low-Income Countries
Mark M. Pitt, Mark R. Rosenzweig, and Md. Nazmul Hassan 1139
- A Social Exchange Approach to Voluntary Cooperation
Heinz Holländer 1157
- The Economic Effects of Production Taxes in a Stochastic Growth Model
Michael Dotsey 1168
- Deposit Insurance, Risk, and Market Power in Banking
Michael C. Keeley 1183
- Overdrafts and the Demand for Money
Avner Bar-Ilan 1201
- Tax Smoothing with Financial Instruments
Henning Bohn 1247

13 CENTRAL

Shorter Papers

Beneficial Concentration	<i>Andrew F. Daughety</i>	1231
Horizontal Mergers: The 50-Percent Bench-Mark	<i>Dan Levin</i>	1238
Input Market Price Discrimination and the Choice of Technology	<i>Patrick DeGraba</i>	1246
Direction of Price Changes in Third-Degree Price Discrimination	<i>Babu Nahata, Krzysztof Ostaszewski, and Prasanna Kumar Sahoo</i>	1254
Third-Degree Price Discrimination and Output: Generalizing a Welfare Result	<i>Marius Schwartz</i>	1259
The Measurement of International Trade Related to Multinational Companies	<i>F. Steb Hipple</i>	1263
The Indirect and Direct Substitution Effects	<i>Masao Ogaki</i>	1271
Auction Institutional Design: Theory and Behavior of Simultaneous Multiple-Unit Generalizations of the Dutch and English Auctions	<i>Kevin A. McCabe, Stephen J. Rassenti, and Vernon L. Smith</i>	1276

Errata

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records	<i>Joshua D. Angrist</i>	1284
On the Optimal Tax Base for Commodity Taxation	<i>John D. Wilson</i>	1287
Presidential Address—The Future of the Income Tax	<i>Joseph J. Pechman</i>	1288

P, 6045

Market Liquidity, Hedging, and Crashes

By GERARD GENNOTTE AND HAYNE LELAND*

In the absence of significant news, hedging strategies were blamed for the stock market crash of October 1987; but traditional models cannot explain how a relatively small amount of selling could cause so large a price drop. We develop a rational expectations model in which prices play an important role in shaping expectations; markets are much less liquid in our model than in traditional models. Discontinuities (or "crashes") can occur even with relatively little hedging. The model is consistent with theories as disparate as Keynes' "beauty contest" insight and Thom's "catastrophe" analysis and suggests means to reduce volatility. (JEL 313)

Immediately following the stock market crash of October 19, 1987, both practitioners and academics sought an explanation based on external events. While several trends were clearly "bad news" for the market, these trends had been revealing themselves over the previous months. It was difficult to isolate new events occurring between October 16 and October 19 that were of sufficient importance to explain the magnitude of the price fall.

The Brady Commission's examination of the October break (Nicholas Brady et al., 1988) therefore centered on internal market causes rather than external events. In particular, the Commission focused attention on a number of large institutions following "price insensitive" strategies such as portfolio insurance.¹

In dramatic language, the Brady Report painted a picture of enormous waves of institutional selling driving down prices ex-

cessively. The report claimed that such sellers suffered from an "illusion of liquidity"; and it buttressed this conclusion by pointing out that a few large sellers alone sold about \$6 billion of stock and stock index futures.

Although there has been a unanimous positive response to the Brady Commission's marshalling of facts, there has not been a unanimous acceptance of their interpretations of these facts. Formal portfolio-insurance strategies were followed by less than 3 percent of stock market funds.² While the \$6 billion sold by portfolio insurers seems a large amount, it represented only 15 percent of total stock and stock index futures volume on October 19. In absolute terms, the \$6 billion amounts to less than 0.2 percent of the roughly \$3.5 trillion of equity value at the beginning of the day. Is it reasonable to think that selling 0.2 percent can drive down prices by over 20 percent—that selling \$6 billion can cause losses of \$700 billion? Of course the answer depends upon the elasticity of demand for stocks. But traditional models imply an elasticity much greater than the market exhibited on October 19, 1987.

*Walter A. Haas School of Business, University of California, Berkeley, CA 94720. The authors thank Fischer Black, Milt Harris, Roy Henriksson, Guy Laroque, Ailsa Roell, Mark Rubinstein, Toshi Shibano, and especially Pete Kyle for extended discussions across many subjects. Support from the Berkeley Program in Finance and from Deutsche Forschungsgemeinschaft, Gottfried-Wilhelm-Liebniz-Förderpreis, during BoWo89 is gratefully acknowledged.

¹Portfolio insurance strategies are dynamic hedging strategies which provide protection by replicating a put option (see Mark Rubinstein and Leland, 1981). These strategies have the property that they tend to sell after the market has declined and to buy after market rises.

²Best estimates suggested \$70–\$100 billion in funds were following formal portfolio insurance programs. On a precrash total stocks value of about \$3.5 trillion, this represents 2–3 percent. Of course, informal hedging strategies such as stop-loss selling may have amounted to considerably more than this (see the survey of Robert Shiller [1987]).

A recent study by Michael Brennan and Eduardo Schwartz (1989) suggests that a 5-percent use of portfolio insurance by investors would have a minimal impact on market prices and volatility. Their model and other commonly studied portfolio/consumption models suggest elasticities of demand for stock far greater than 1—more than 100 times the elasticity implied by the Brady Commission's conclusions. Information changes, rather than selling by portfolio insurers, are needed to explain October 19 in these standard models.

Other evidence does not seem to confirm a strong connection between the crash and portfolio insurance. If short-run selling were the cause of the decline, we might expect a quick reversal, but this did not occur. Furthermore, Richard Roll's (1989) cross-market studies showed little correlation between October 1987 performance and various aspects of markets, including whether portfolio insurance was used.

In sum, the crash of 1987 presents the following dilemma to current financial models: the amount of selling seems insufficient to explain the large price drop observed on October 19. Thus, information changes seem necessary to explain the drop, but no such information changes can be documented.

Parallels with the crash of 1929 may be useful in understanding the crash of 1987. Like 1987, no significant economic news was associated with the period immediately surrounding the earlier crash. Several large declines preceded the crash of 1929—as they did in 1987. Volatility increased markedly in the weeks preceding both drops. In both cases, hedging strategies were discussed as a possible contributing factor: stop-loss orders in 1929 and portfolio insurance in 1987. In 1929, stop-loss strategies were used for portfolio protection but were also triggered by margin calls, which led to greater controls over margined stock-buying following the crash.

Because of these similarities, we would hope that an explanation of the 1987 crash would also be relevant to the 1929 crash. The explanation cannot be entirely in terms of futures markets and portfolio insurance, since neither existed in 1929.

In this paper, we develop an explanation of market "crashes" that reconciles the strands of evidence above. This explanation is not based on important changes in information, and therefore it is consistent with the failure to observe any significant events that directly "caused" the 1987 (or 1929) decline. In fact, we define a "crash" to be a discontinuity in the relationship between the underlying environment and stock prices: an infinitesimal shift in information (or other small shock) can lead to a major change in stock market level.³

Our explanation of crashes is based on unobserved plans of investors to hedge against losses. In 1929, stop-loss strategies were used. In 1987, portfolio insurance and stop-loss strategies were followed. We develop a "price pressure" argument akin to that of Sanford Grossman (1988a). However, this argument must meet two criticisms:

- (1) How can relatively small amounts of hedging drive down prices significantly?
- (2) Why didn't stock prices rebound the moment such selling pressure stopped?

Our model answers the first question by examining the determinants of market liquidity. An important aspect of financial markets is that only a small proportion of investors actively gather information on future economic prospects or asset supply. Other investors look to current prices to impute information about future prices. This dual role of prices—affecting demand both through the budget constraint and through expectations—leads to very different price elasticities than traditional models, in which prices play only the first role. Only recently have financial economists begun to explore

³Such discontinuities are commonly observed in physical systems and have been the recent subject of study by mathematicians examining "catastrophe theory." In a preliminary paper (Gennotte and Leland, 1987), we considered a simple model of stock market discontinuities.

the implications for markets in which prices play both roles.⁴

In such environments, there is an important difference between observed and unobserved supply changes. If there are relatively few informed investors, markets may be much less liquid (and therefore more fragile) than traditional models predict when unobserved supply shocks occur.⁵ We show that relatively small unobserved supply shocks can have pronounced effects—more than 100 times greater than the effects of observed supply shocks—on current market prices.

Unobserved supply shocks have greater price impact as a consequence of investors inferring information from prices. A supply shock leads to lower prices, which in turn (since the shock is unobserved) leads uninformed investors to revise downwards their expectations. This limits these investors' willingness to absorb the extra supply and causes a magnified price response.⁶

Our model answers the second question by showing how a discontinuity in market prices can occur if hedging plans generate very large trades. Hedging plans create additional supply as price falls. A small information change can trigger lower prices, which, because of hedging, lead to greater excess supply and a further fall in prices. Thus, a small change in information can lead to a dramatic fall in prices, with no immediate rebound occurring. This feature of crashes distinguishes our results from

those of Grossman (1988a,b). We demonstrate that such a "meltdown" scenario obtains only for an unrealistically large hedging activity when the hedging trades are perfectly anticipated. If, however, investors (or a fraction of investors) are unaware of hedging plans, crashes can occur for much smaller levels of hedging activity. The discontinuities arise because investors are unable to perfectly distinguish hedging activity from information-based trades and therefore adjust downward their expectations of future prices. Imperfect anticipation of hedging activity relaxes the rational expectation requirement, but in a realistic fashion. Moreover, our estimate of hedging plans in place in 1987 approximates the threshold at which discontinuities occur in the imperfect anticipation case, providing a potential explanation for the 1987 crash.

Finally, our model suggests that some changes in market organization can radically reduce the likelihood of crashes. The most important such change is increasing market knowledge about the size and trading requirements of hedging programs. Preannouncement of trading requirements can lessen the impact of such trades by a factor greater than 100. To the extent that the specialist's book helps reveal the nature of order flow, this information should be made available to all traders. As suggested by Grossman, the use of put options to implement hedging may also serve to smooth markets.

The model builds from the work of Sanford Grossman and Joseph Stiglitz (1980), Martin Hellwig (1980), Douglas Diamond and Robert Verrecchia (1981), Anat Admati (1985), Gennotte (1985), and Albert Kyle (1985). We postulate a subset of informed investors who receive noisy signals about future market values. Random supply keeps these investors from perfectly inferring the aggregate information from price. Some investors, however, whom we characterize as market-makers, receive information about the size of the random supply. This information enables market-makers to distinguish, at least partially, selling that is information-based from selling that is motivated by liquidity considerations. We show

⁴See Grossman (1976), Grossman and Joseph Stiglitz (1980), Martin Hellwig (1980), Douglas Diamond and Robert Verrecchia (1981), Albert Kyle (1985), and Anat Admati (1985). These papers focus on the role of prices in aggregating information.

⁵This point is discussed informally in Leland and Rubinstein (1988) and in D. Cutler, James Poterba, and Lawrence Summers (1989). Fischer Black (1988) describes a model in which shocks to expectations—rather than supply—can cause large price changes.

⁶Such models reflect a rational expectations view of Keynes' famous "beauty contest" metaphor, that successful investors must base their investments on their expectations of others' expectations of value, rather than solely on their own estimates of value. Thus price, reflecting others' expectations, rationally conditions each individual investor's expectations, and bandwagon or "herd" effects can result.

that market-makers provide a significant source of liquidity in meeting random supply shocks.

We also allow for the presence of hedging programs such as stop-loss orders and portfolio insurance. These hedging strategies are usually nonlinear functions of the equilibrium price; hence, the resulting rational expectations equilibrium price is, in general, a nonlinear function of the signals. We examine the effects of these strategies on market equilibrium and stability for alternative specifications about the observability of hedging. The possibility of price discontinuities, with the implication of crashes, is new to our model.

I. Informed and Uninformed Investors

We assume that investors may be informed or uninformed. The informed investors can be subdivided into two types, who differ in terms of the signals they are able to observe. Thus, in all there are three classes of investors:

- (1) *uninformed* investors (denoted U), who observe only the equilibrium price p_0 ;
- (2) *price-informed* investors (I), who observe a personal, unbiased signal p_i^* on future price (or liquidation value) p and also observe p_0 ;
- (3) *supply-informed* investors (SI), who observe a common supply signal S and the equilibrium price p_0 .⁷

The price-informed investors can be thought of as having (personal) information about economic fundamentals which are noisy predictors of future price. Supply-informed investors can be thought of as market-makers who have information about the sources of order flows: the size of new issues, portfolio restructurings, and other elements of liquidity trading.⁸

Our model allows arbitrary proportions of investors in each class. The relative proportion of investors who are informed versus uninformed is a key determinant of market liquidity. Informed investors, particularly supply-informed investors, will absorb a substantial proportion of liquidity-trading demands. Even when they are relatively few, informed investors constitute an important fraction of the supply of liquidity.

Thus, an important empirical question is the relative number of investors of each type. While data on this question are difficult to gather, we do have some evidence that informed traders, particularly supply-informed traders, are relatively small as a fraction of total market capital.

Among supply-informed investors are specialists and other market-makers (including "upstairs" desks) who adjust their positions in response to changing demand for liquidity. Because of their role as market-makers, these investors have special information on the nature of demand. Through the order book or simply on the basis of their knowledge of institutional trading, market-makers can learn (perhaps imperfectly) about the volume of noninformation (or "liquidity") trading versus trading based on information.

The funds committed to supply-information gathering (and providing liquidity) depend upon the return to this activity. In some cases, competitive forces will determine the amount provided. In other cases, institutional factors such as the specialist system may limit the number of potential entrants. We shall see below that such limitations can importantly affect the stability of markets.

There is no way to provide an exact quantification of market-making capital. However, it clearly is small relative to the \$3.5 trillion of equity investment. For example, the total capital of New York Stock Ex-

⁷Investors who are both price-informed and supply-informed can easily be incorporated in this framework. Adding an I/SI investor has the same impact as adding an I and an SI investor separately.

⁸Our "market-makers" behave competitively, taking prices as given. In contrast with Lawrence Glosten and

Paul Milgrom (1985), we do not require that all trades be completed through the market-makers, or that market-makers are risk-neutral. Our market-makers absorb the aggregate excess supply generated by other traders at the equilibrium price. We discuss their contribution to market liquidity in Section IV.

change specialists, including lines of credit, is approximately \$3 billion (Brady et al., 1988 p. VI-40). Total capital committed by "upstairs" trading houses and other forms of market-making may be four or five times this number;⁹ but at \$15–\$20 billion, these supply-informed funds would represent about 0.5 percent of total market capital.

Capital devoted to "price-informed" market timing is even more difficult to estimate. It would include funds that explicitly gather information about future economic prospects ("fundamentals") and engage in market timing strategies reflecting this information. While many funds actively alter their exposures to individual stocks, most do not actively alter their total stock exposure based on information about future economic trends, perhaps because long-term success stories have been so rare (see Roy Henriksen, 1984); but a few do. The single most prominent market timing strategy is "tactical asset allocation," utilized by perhaps \$20 billion of assets. A total of \$70 billion, or about 2 percent, might be a guess for price-informed funds which actively gather information about future prices and trade on it.¹⁰

This leaves most investors in the class we term "uninformed." "Passive" might be a somewhat less pejorative description of these investors, who participate in the market "for the long haul" and do not move in and out based on information about fundamentals or current liquidity trading. The relative lack of popularity of information-based market timing strategies suggests that most investors belong to this class.

II. Market Equilibrium

A single risky security is traded. Its future price (or liquidation value) p is a normally

distributed random variable with unconditional variance Σ and unconditional expectation \bar{p} . All investors share this prior distribution of future price. Current price is determined by supply and demand. Riskless bonds are also traded, and the riskless rate is zero.¹¹

A. Supply

The net supply of stocks is a fixed amount \bar{m} , modified by two additional factors:

- (1) A random and exogenously determined net supply created by "liquidity traders." This shock is composed of two pieces: an unobserved liquidity shock, L , distributed $N(0, \Sigma_L)$; and a liquidity shock, S , distributed $N(0, \Sigma_S)$, which is observed by all supply-informed investors.¹² S and L are assumed to be independently distributed.
- (2) A deterministic supply by hedgers, rebalancers, and others who utilize dynamic strategies akin to portfolio insurance. The supply of stock from these strategies is a known function of the current market price p_0 . We denote this supply by π , a decreasing differentiable function of p_0 . This hedging demand will be observed by different market participants in the alternative environments that are considered below.

Thus, the total supply is

$$\bar{m} + L + S + \pi$$

or

$$(1) \quad m + \pi$$

where $m = \bar{m} + L + S$.

⁹Conversations with officers at major investment banking firms.

¹⁰We assume that the informed investors have information of value. It is difficult to assess the fraction of the market timing based on "quality" fundamental information, but it is most probably smaller than the fraction engaged in timing in general. Investors trading on spurious information would add to the amount of random "liquidity" trading.

¹¹All the results extend in a straightforward way to the case of nonzero interest rates.

¹²An equivalent formulation would allow the supply-informed investors to receive a noisy signal about total liquidity trading and not distinguish S from L . A simple transformation of variables allows the alternative interpretation.

B. Demand

As discussed above, there are three classes j of investors characterized by the information signals (if any) they receive. All investors maximize expected utility of terminal wealth over a single period. Preferences are assumed to exhibit constant absolute risk aversion. The utility function of each investor in class j is a function of terminal wealth W and is given by

$$U_j(W) = -\exp(-W/a_j).$$

Expectations depend upon the signals that investors observe. Each price-informed investor i observes $p_i^t = p + \varepsilon_i$, where p is the true future price and ε_i denotes a noise term uncorrelated with other random variables and uncorrelated across price-informed investors. Both p and ε_i are assumed to be normally distributed, as $N(\bar{p}, \Sigma)$ and $N(0, \Sigma_\varepsilon)$, respectively.¹³ Supply-informed (SI) investors receive a common signal, S , where S is the liquidity supply, distributed $N(0, \Sigma_S)$ and observed only by SI investors. All investors can observe the current market price and use this (and their other information) to determine their conditional distributions of future price.

All investors in class j have the same conditional variance Z_j for future price. The expectation of the future price conditional on the information available to an agent i belonging to class j is denoted \bar{p}_j^i . It is well known that, given exponential utility and normality, the portfolio optimization problem leads to a demand for shares by investor i in class j equal to

$$n_j^i = a_j Z_j^{-1} (\bar{p}_j^i - p_0).$$

There are w_j investors of type j . Demand per investor in class j , $n_j \equiv \sum_i n_j^i / w_j$, is equal to

$$n_j = a_j Z_j^{-1} (\bar{p}_j - p_0)$$

¹³ For simplicity and notational convenience, we assume that the ε_i are i.i.d. across agents. Differences in information precision would not affect our results.

where \bar{p}_j is the mean expected future price for investors in class j . All supply-informed investors observe the same signals and hence have the same conditional expected price. This also is the case for uninformed investors. Price-informed investors receive different signals, ε_i . However, as their number increases, the mean expected future price converges to the actual future price p by the law of large numbers.¹⁴

The relative market power of investor class j , k_j , is defined as the ratio of the weighted risk tolerance of the class to the sum of the weighted tolerances:

$$k_j \equiv a_j w_j / \sum_j a_j w_j.$$

Define the normalized total demand D as the sum of the individual classes' demands divided by the weighted sum of the three classes' risk tolerance, $\sum_j w_j a_j$:

$$(2) \quad D \equiv \frac{\sum_j w_j n_j}{\sum_j w_j a_j} = \sum_j k_j Z_j^{-1} (\bar{p}_j - p_0)$$

Similarly, the supply parameters: π , \bar{m} , S , and L are normalized by dividing the original parameters by $\sum_j w_j a_j$; we do, however, keep the same notation. Our analysis focuses on relative proportions and is thus unaffected by this normalization.

C. Equilibrium

Equilibrium of supply (1) and demand (2) yields the equation for equilibrium price:

$$Z^{-1} p_0 + \pi(p_0) = \sum_j k_j Z_j^{-1} \bar{p}_j - m$$

¹⁴ Hellwig (1980) shows the error terms ε_i do cancel in the limit of a sequence of finite economies where the relative proportion of investors in each class remains fixed and where the total number of investors and the supply parameters grow without bound at the same rate. Individual agents are thus price-takers and, importantly, the individual error terms ε_i do not affect prices.

where Z^{-1} is $\sum_j k_j Z_j^{-1}$. We may now characterize the equilibrium price function relating current price to future price (as revealed by the average of individual signals p_i), unobserved liquidity supply (L), and observed supply (S).

In our postulated environment, with all investors cognizant of hedging supply $\pi(p_0)$, we can show the following.

THEOREM 1: *There exists a rational expectations equilibrium (REE) of the form*

$$(3) \quad p_0 = f(p - \bar{p} - HL - IS)$$

where $f(\cdot)$ is a correspondence and where H and I are real constants that depend only on the agents' relative market power and on the means and variances of the random variables.

The proof and all derivations, as well as expressions for $f^{-1}(\cdot)$, H , and I , are provided in the Appendix.

The price correspondence $f(\cdot)$ in Theorem 1 can be discontinuous and multivalued. We can, however, characterize precisely the situations in which crashes can be ruled out as follows.

PROPOSITION 1: *$f(\cdot)$ is a continuous function if and only if $Z^{-1}p_0 + \pi(p_0)$ is strictly monotonic in p_0 .*

PROPOSITION 2: *In the absence of hedgers [$\pi(p_0) = 0$], $f(\cdot)$ is a continuous function, and no "crashes" can occur.*

In the absence of hedgers or if the hedging supply is a linear function of equilibrium prices, f is a linear function [equation (A4) in the Appendix]. The REE function f is nonlinear if the hedging supply is a nonlinear function of equilibrium prices p_0 . Even though current price p_0 is not normally distributed due to the nonlinearity of $f(\cdot)$, investors recognize that a simple transformation of p_0 , namely $f^{-1}(p_0)$, is normally distributed and can be used to condition expectations of future price p . In the following two sections, we examine aspects of equilibrium price behavior in the absence of hedgers.

III. The Nature of Equilibrium Pricing: An Example

Consistent with our earlier discussion, we consider an example in which there are relatively few price-informed and supply-informed investors. We assume that 0.5 percent of investors (market-makers) are supply-informed and that 2 percent of investors are price-informed. There is no hedging supply π ; this will be introduced in Section IV.

Several other parameters must be specified before the model is complete. A key parameter is the quality of the information signal received by the price-informed investors. The better the signal, as expressed by the signal-to-noise ratio, the lower the conditional variance Z_1 for the price-informed investors.

We assume that the quality of the signal received by each price-informed investor is not very high. Specifically, we assume that the price-informed investors' signal-to-noise ratio is 0.2. Thus, if Σ is the *ex ante* variance of future price and Σ_e is the variance of each price-informed investors' future price signal, then Σ_e is five times Σ . This assumption implies that (in equilibrium) the price-informed investors' conditional standard deviation for future price is 19.1 percent, rather than the 20 percent of uninformed investors, who observe price only. This slight improvement seems consistent with the perceived difficulty in predicting future market prices.¹⁵

Also important is the fraction of total liquidity-supply shock that, on average, can be observed by supply-informed investors. Since S is observed and L is not, this fraction can be parameterized by the ratio Σ_S/Σ_L . If the ratio is high, then supply-informed investors on average will observe most of the total liquidity shock. We assume that the ratio is 1: On average, supply-

¹⁵ While each individual signal about future price is quite noisy, the average signal perfectly reflects future price, as in Hellwig (1980). But individual investors cannot "back out" the true future price from current price, because supply is noisy.

informed investors receive a signal that reveals information about half the total liquidity-supply shock. Conditional on the supply signal, a supply-informed investor estimates a 19.2-percent standard deviation for future price.

Our example is consistent with a linear rational expectations price function as in Theorem 1. We choose Σ , the *ex ante* variance of future price, and \bar{m} , the fixed supply, to equate the expected return on the risky security to 6 percent and the standard deviation (conditional on current price) to 20 percent.¹⁶ Finally, we find a variance of supply Σ_m such that the variance of p_0 , the current price, is equal to the variance of the future price p , conditional on p_0 . This provides the example with intertemporal consistency: expected future price volatility (given current price) equals current price volatility. Parameters for the example are summarized in the Appendix.

The rational expectations equilibrium price function (3) for our example is

$$(4) \quad p_0 = 0.5(p - \bar{p} - 19.95L - 8.14S) + 1.$$

Given this price function and the volatilities of future price and liquidity surprises, the standard deviation of p_0 is 20 percent, as is the standard deviation of p conditional on p_0 . The example is chosen to reflect "reasonable" parameters when the model's single risky security is interpreted as the stock market portfolio and will be used in subsequent sections to illustrate aspects of market behavior.

IV. Stock Market Liquidity

Because of the Brady Commission's focus on limited stock market liquidity, we examine the impact of changes in supply on market price. We postulate a small percent-

age change in supply and determine the resulting percentage change in the equilibrium price from the pricing relation (3). This will then determine the (inverse of the) price elasticity of the market. We interpret greater price elasticity as a more liquid market. In this section, we continue to assume that there is zero net hedging: $\pi = 0$.

We study three possibilities: supply increases, and

- (1) the increase in supply is known to all investors;
- (2) the supply-informed investors (only) receive an accurate signal about the increase in supply;
- (3) no signal is received by the supply-informed investors (or anyone else).

The first possibility is modeled by letting the expected supply \bar{m} change. A change in expected supply will be common knowledge and will not affect investors' expected future price. From equations (3) and (A4) in the Appendix,

$$\text{Elasticity} = - \frac{\frac{\delta \bar{m}}{\bar{m}}}{\frac{\delta p_0}{p_0}} = \frac{p_0}{\bar{m}} Z^{-1}.$$

Given the example (4), we find an elasticity of 17: a 1-percent observed supply increase will lead to a 0.06-percent fall in price. Such a high elasticity is very much in line with the predictions of traditional models, which do not postulate that investors learn from market prices. Investor classes participate proportionately to their market power k_j in absorbing the increase in supply.

The second possibility is modeled by a small increase in the random supply S which is observed only by the supply-informed investors. In this case,

$$\text{Elasticity} = - \frac{\frac{\delta s}{\bar{m}}}{\frac{\delta p_0}{p_0}} = \frac{p_0}{\bar{m}} \frac{1}{FI}$$

¹⁶Recall that the interest rate is normalized to 0. Thus, the assumed return of 6 percent represents a 6-percent premium over the riskless interest rate. The risk and excess return of the risky security in our example are consistent with the long-term risk and excess return of the stock market as estimated by Ibbotson Associates (1985).

where F is the slope of the price function f , which is linear in our example.

In the example (4), we find an elasticity of 0.16: a 1-percent partially observed increase in supply will lower price by 6 percent. Remarkably, this is only 1 percent of the elasticity above. This is because investors, with the exception of the supply-informed investors, revise downward their expectations (which are conditional on p_0) as price falls. Thus, they are less willing to absorb the increased supply. Indeed, in our example, the supply-informed investors absorb about 54 percent of any increase in liquidity supply, which they observe, even though they constitute only 0.5 percent of investors.¹⁷

Price-informed traders, who represent 2 percent of investors, absorb another 18 percent of the increase in liquidity supply. They are more willing to buy as prices fall because (on average) they receive signals about future price that moderate the fall of expected future price. Uninformed investors have no such signals and can only infer from current price. Because they impute a lower future price as current price falls, they absorb but 28 percent of the increased supply, despite the fact that they constitute 97.5 percent of investors.

The final possibility is modeled by an increase in L . In this case, elasticity falls even further, since the supply-informed traders will not observe the increase in supply and will not increase their demand. From (3), we can determine that

$$\text{Elasticity} = - \frac{\frac{\delta L}{\bar{m}}}{\frac{\delta p_0}{p_0}} = \frac{p_0}{\bar{m}} \frac{1}{FH}.$$

¹⁷The exponential utility model does not limit purchases by investors to their initial wealth. If we imposed a "no leverage" condition, elasticity in this case would be even lower. This is because in our equilibrium, supply-informed investors will buy tremendous amounts of stock (on a per capita basis) when prices fall. This would only be possible if they can undertake levered stock positions.

In the example (4), elasticity will be 0.07, or about 1/250 of the elasticity predicted by traditional portfolio/consumption models. A 1-percent unobserved increase in supply will lower prices by 14 percent. In this case, price-informed investors will absorb about 40 percent of the supply increase, and uninformed investors absorb the remaining 60 percent.¹⁸ But price must fall precipitously to induce them to absorb the extra supply.

The model therefore resolves the paradox of low versus high demand elasticity. If supply changes are unobserved, all investors will revise downward their expected future price and will absorb the increased supplies only after price has fallen substantially. Price-informed investors will have somewhat greater elasticity of demand than uninformed investors, since they receive independent information about future prices. However, their contribution will be minimal if they are few, or if their price information is very noisy.

How supply-informed investors, or market-makers, contribute to market liquidity (and therefore to price volatility) depends on the quality of the supply signal they observe. When their signal has low precision, they side with the uninformed investors, who always sell when prices rise and always buy when prices fall. Their average selling price therefore is higher than their average buying price, providing a positive spread much like the uninformed market-maker in Lawrence Glosten and Paul Milgrom (1985). Because their actions are not aggressive and their numbers are relatively small, market-makers with poor supply information will reduce volatility only marginally.

When market-makers receive more precise information about the extent of liquidity trading, their behavior changes. They aggressively take the other side of observed liquidity trades, thereby reducing the price volatility associated with liquidity shocks;

¹⁸Supply-informed investors play little role in this scenario, since they observe $S = 0$. In fact, they actually sell (a small amount) rather than buy, since observing $S = 0$ implies that price information is more likely to be negative, given the fall in prices.

but supply-informed investors will trade in the same direction as informed traders when their information indicates that there is little liquidity trading. This behavior tends to accentuate price moves related to information and suggests that a further examination of the market-makers' role is warranted.¹⁹

Only the supply-informed market-makers will have a high elasticity of demand and the consequent ability to absorb liquidity trades. But these participants are relatively few in number and risk aversion ultimately limits their ability to absorb liquidity selling without a substantial price drop. In sum, traditional models will grossly overestimate the liquidity of financial markets, unless all investors observe the increase in supply.

V. Hedging Strategies and Market Stability

We now consider the situation when there is hedging activity in the market. Hedgers sell as stock prices fall. They do so to protect themselves against further potential losses. Whether hedge programs are carried out by portfolio insurance programs or by less formal means such as stop-loss orders, the result is the same: supply increases as prices fall. This selling must be absorbed by other investors.

It is generally believed that hedge programs can make markets more volatile. But can they lead to a crash or "meltdown," in which selling begets selling and prices plunge without stop? Phrased more formally, can the function relating price to underlying information become discontinuous?

We examine these questions by adding a hedging supply (π) to a market that previously did not have such a supply. At the initial equilibrium price ($p_0 = 1$), we normalize hedging supply to be zero. For prices below $p_0 = 1$, the hedging supply will be positive; for prices above, negative. We assume that the hedging supply is a continuous and differentiable function of p_0 .

¹⁹Perhaps recognizing the incentives of supply-informed market-makers to accentuate price movements by their own trading activities, exchange rules require specialists to maintain "orderly markets."

The key to the stability of markets is the extent to which hedging strategies are observed by investors. Parallel to our discussion of market liquidity, we consider three cases: all investors observe the hedging supply function π ; only supply-informed investors observe hedging; and no investors observe hedging. In the first case, agents have perfect knowledge of the market structure; in the others, some agents underestimate hedging activity.²⁰

Theorem 1 characterized equilibrium when all investors can observe the hedging supply function $\pi(\cdot)$. We can further show the following.

PROPOSITION 3: *When all investors observe hedging supply π , (i) the current equilibrium price will be more volatile than when there is no hedging supply, and (ii) the equilibrium price function can be discontinuous when hedging supply is sufficiently large.*

Although discontinuities and, therefore, crashes can occur with full observability of π , we shall show that crashes are unlikely in this environment.

We now characterize equilibrium when hedging is partially observed or unobserved.

THEOREM 2: *If hedging supply $\pi(\cdot)$ can be observed only by supply-informed investors or is totally unobserved, there exists a price equilibrium*

$$p_0 = f(p - \bar{p} - HL - IS)$$

²⁰Through time, investors might learn of the existence of hedgers. Modeling this would require a multi-period framework. However, the amount of hedge trading typically is small if the price level is not "too" close to the critical points, making it difficult to infer the extent of hedging. Further, when hedgers first enter the market, there is no mechanism whereby uninformed traders could immediately recognize their presence. An extension of our model might allow investors to have some prior probability distribution of hedging supply given current price and use observed prices to infer the likelihood of hedging activity as well as liquidity shocks and future price information. We hope to explore this more complex problem in subsequent work.

where $f(\cdot)$ is a correspondence that depends upon the extent of observability, but H and I are real constants which are identical to those when hedging supply is fully observed.

Partial observability or nonobservability therefore does not affect the argument of $f(\cdot)$ but does affect its functional form. In the context of Theorem 2, some investors do not observe hedging supply and thus make an erroneous assumption on the functional form for equilibrium prices. This incorrect assumption is reflected in the functional form of actual equilibrium prices. The agents who are aware of hedging plans, however, know the actual functional form. For the cases in which some agents are aware of hedging plans, the same equilibrium would obtain if the agents were unaware of hedging plans but were able to identify the hedging trades as liquidity (i.e., information-free) trades. This property explains why the difference in the excess demand equations (5) below amounts to adding hedging activity π to the expected supply \bar{m} [in (5i)], to the observed liquidity trades S [in (5ii)], and to the unobserved liquidity trades L [in (5iii)].

We further show the following.

PROPOSITION 4: *When only supply-informed investors observe hedging activity π , (i) the current price will be more volatile than when all investors observe a similar amount of hedging activity, and (ii) the equilibrium price can become discontinuous at a lower level of hedging activity than when all investors observe hedging.*

PROPOSITION 5: *When all investors are ignorant of hedging activity π , (i) current price is even more volatile than when only supply-informed investors observe hedging activity, and (ii) discontinuities can occur at even lower levels of hedging activity.*

The maximal hedging level before price discontinuities or "crashes" can occur thus depends critically on whether hedging is observed. It also depends upon the nature of hedging strategies. We assume that a fraction ω of assets are protected by a put-

option replicating strategy.²¹ The supply created by this portfolio-insurance hedging strategy will depend on the current stock price p_0 . The incremental hedging supply when future price is p_0 , relative to the supply at the initial equilibrium price ($p_0 = 1$), is given by

$$\pi = \omega \{ N[d_1(1)] - N[d_1(p_0)] \}$$

where ω is the fraction of assets subject to the hedging strategy, $N(\cdot)$ is the cumulative normal distribution function, and d_1 is given by the Black-Scholes formula

$$d_1(p_0) = \frac{\ln\left(\frac{p_0}{K}\right) + \frac{1}{2}\sigma^2}{\sigma}$$

where K is the striking price of the option and σ is the standard deviation of p conditional on p_0 .²² Note that $\pi'(p_0)$, the derivative of the hedging supply with respect to p_0 , is negative for large p_0 and becomes more negative as p_0 falls, before eventually approaching zero as p_0 falls to zero.

With the three alternative specifications above for $f(\cdot)$ depending on observability, we can derive the excess-demand functions as p_0 varies. From the Appendix, the equations for excess demand are given by

$$(5i) \quad XD_A = \frac{1}{H} \left[p - \bar{p} - HL - IS + \frac{Z^{-1}(\bar{p} - p_0) - (\bar{m} + \pi)}{Z^{-1} - \Sigma^{-1}} \right]$$

²¹Put-replicating strategies are just one possible type of hedging. Others might include stop-loss, "constant proportion of surplus" policies, or do-it-yourself strategies. We examine put-option replication because it was the most prevalent of formal protection strategies on October 19.

²²See Black and Myron Scholes (1973). Our formula assumes that the interest rate has been normalized to 0, and assumes a one-year time horizon. Note that the Black-Scholes hedge replicates a put option when future price p follows a lognormal process; our model presumes that p is normally distributed.

when all investors observe π ,

$$(5ii) \quad XD_P = \frac{1}{H} \left[p - \bar{p} - HL - IS + \frac{Z^{-1}(\bar{p} - p_0) - \bar{m}}{Z^{-1} - \Sigma^{-1}} \right] - \frac{I}{H} \pi$$

when only supply-informed investors observe π , and

$$(5iii) \quad XD_U = \frac{1}{H} \left[p - \bar{p} - HL - IS + \frac{Z^{-1}(\bar{p} - p_0) - \bar{m}}{Z^{-1} - \Sigma^{-1}} \right] - \pi$$

when no investors observe π .

These three functions are graphed in Figure 1 for the parameters in our earlier example, with 5 percent of investors following a put replicating hedge strategy with the protected level being 90 percent of initial price ($\omega = 0.05$, $K = 0.9$). Note that the fully anticipated excess-demand function is the flattest; neither it nor the partially anticipated excess-demand function is "backward bending." However, the unanticipated excess demand function is backward-bending. The three curves in Figure 1 intersect at $p_0 = 1$. Thus, in the absence of future price or liquidity shocks, the price $p_0 = 1$ is an equilibrium in all three cases.

Now, suppose that information signals about future price become slightly more pessimistic: $p - \bar{p} = -0.01$, a 1-percent downward shock. This will cause demand to fall slightly, thereby shifting all three curves to the left by the same small amount. Figure 2 depicts this shift.

To restore equilibrium, price will fall in all three cases, until excess demand again is zero. The excess-demand curve when hedging is unobserved has the steepest slope: the resulting price drop (2.7 percent) to restore equilibrium will be greater than the price drop (0.7 percent) to restore equilibrium in the partially observable case, which in turn will be greater than the price drop (0.5 percent) to restore equilibrium in the fully observable case.

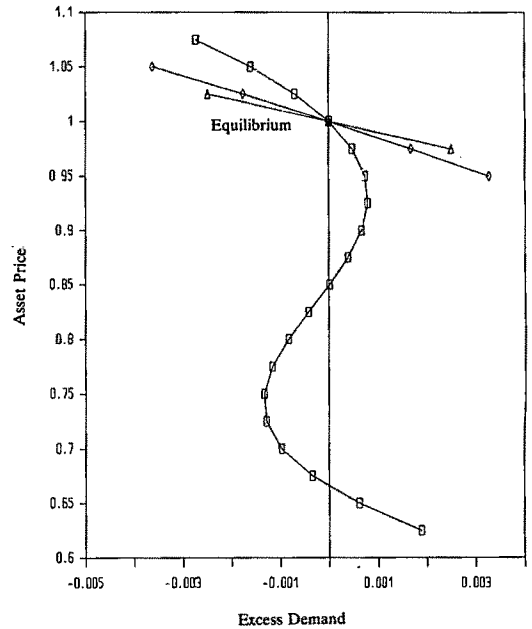


FIGURE 1. AGGREGATE EXCESS DEMAND IN THE ABSENCE OF FUTURE PRICE OR LIQUIDITY SHOCK. SYMBOLS: \square = UNOBSERVED; \diamond = PARTIALLY OBSERVED; \triangle = ANTICIPATED

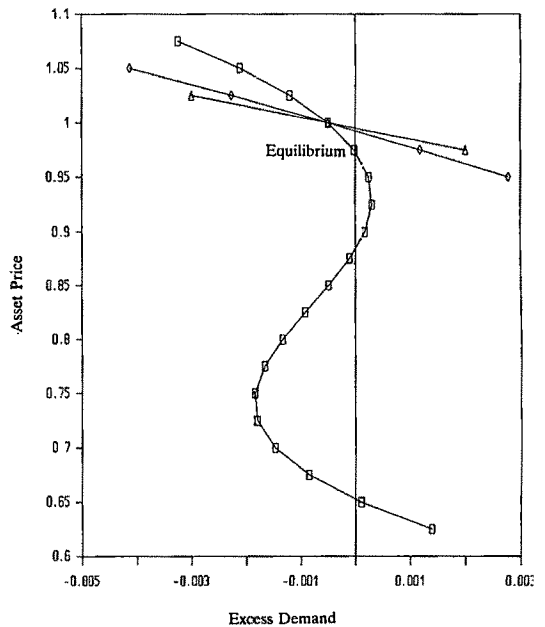


FIGURE 2. AGGREGATE EXCESS DEMAND WITH 1-PERCENT DOWNWARD SHOCK. SYMBOLS: \square = UNOBSERVED; \diamond = PARTIALLY OBSERVED; \triangle = ANTICIPATED

In short, the market price is more volatile in response to future price shocks when hedging supply is unobservable. It will also be greater in all cases if ω , the proportion of hedgers, becomes larger. When hedging activity is unobserved, volatility increases because investors believe a change in fundamentals is more probable, creating a magnified price response.

A. Prelude to a Crash

We continue to examine market behavior as information about future price becomes (continuously) more pessimistic. Figure 3 indicates the situation for $p - \bar{p} = -0.016$. Relative to our initial equilibrium (at $p_0 = 1$), the average of future price signals is now 1.6 percent more pessimistic than the case in Figure 2. Of course, price must drop further to restore equilibrium. This in turn creates further portfolio insurance selling.

How far does price drop? This depends on how completely portfolio-insurance selling is observed. If every investor observes π , the price falls from 1 to 0.992, or 0.8 percent; if it is observed only by the supply-informed, price falls by 1.2 percent; but if no one can observe π , price will fall 7.25 percent in response to the signal (-0.016), almost t n times as far as when π is fully observed.

Note that, in the case of unobserved π , the market also becomes more sensitive to future price signals. This can be seen in Figures 2 and 3 by noting the fact that excess demand is becoming a steeper function of p_0 . Thus, volatility of the market is increasing as p_0 falls.

The move from the situation in Figure 1 to the situation in Figure 3 seems to correspond to the steady erosion of confidence that occurred during the month leading up to October 19, 1987. As the Brady Commission Report documented, a number of negative economic trends came to light during this period: interest rates were rising; the dollar was weakening; tensions in the Middle East were increasing; and so on. In our model, this is reflected by a sequence of negative signals about future price.

As the market fell, portfolio-insurance programs became more active. At higher

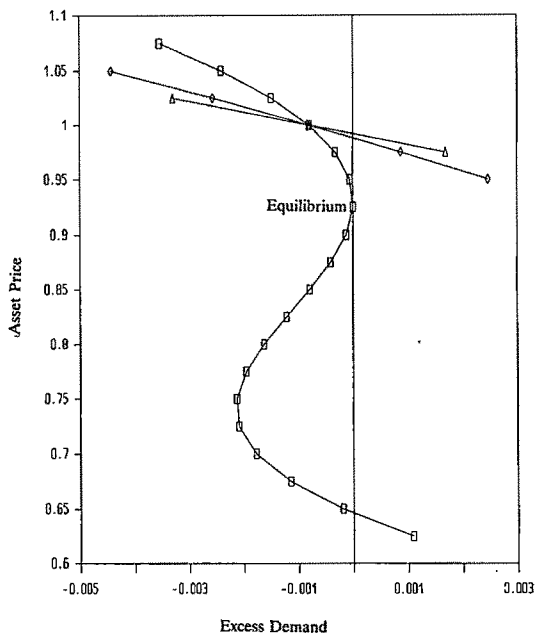


FIGURE 3. AGGREGATE EXCESS DEMAND WITH 1.6-PERCENT DOWNWARD SHOCK. SYMBOLS: \square = UNOBSERVED; \diamond = PARTIALLY OBSERVED; \triangle = ANTICIPATED

market levels, not much hedging was necessary, given the relatively low levels of protection chosen by many pension funds. As the market fell closer to the desired protection level, greater hedging was needed, and the market became more volatile. Yet although portfolio insurance was beginning to attract some public attention, it was largely unknown to the majority of investors and not fully understood even by market professionals. It was Friday, October 16, 1987.

B. The Crash

Figure 3 shows the market at a critical point when hedging strategies are not observed (such as on October 16). Prior to October 16, 1987, prices had fallen strongly over the previous several trading sessions, with great volatility. Over the weekend, a bit more negative news came into the market—nothing earthshaking, but enough to shift the backward-bending excess-demand curve a fractional amount further to the left.

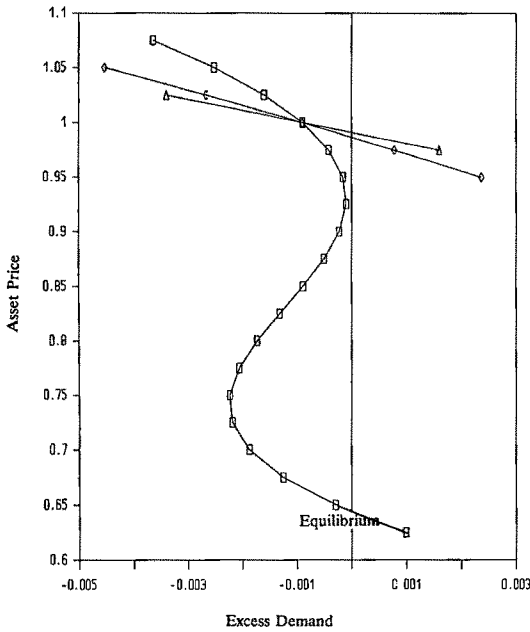


FIGURE 4. AGGREGATE EXCESS DEMAND WITH 1.8-PERCENT DOWNWARD SHOCK. SYMBOLS: \square = UNOBSERVED; \diamond = PARTIALLY OBSERVED; \triangle = ANTICIPATED

Figure 4 illustrates the situation as it may have been on Monday morning, October 19, 1987. The value of $p - \bar{p}$ has fallen to -0.018 , slightly below its previous level. The marginally negative news over the weekend, coupled with further portfolio-insurance selling (including some resulting from Friday's decline) led to rapidly falling prices.

Observing these falling prices, uninformed investors (rationally) concluded that highly negative information must have been received by the price-informed investors. (Indeed, the following day's newspapers vainly sought the information event which "must" have triggered the crash.) As reported by Robert Shiller (1987), the majority of investors stood on the sidelines or bought only limited amounts, consistent with a conviction that something unknown but terrible must have happened. Investors surveyed by Shiller reacted more to the crash itself than to outside news. Meanwhile, hedgers were selling ever larger amounts.

As Figure 4 shows, excess supply actually increased as prices began to fall, leading

them to fall even further. The feared meltdown was actually happening. Only when hedgers had largely completed their selling did the market stabilize, but at a much lower level. In Figure 4, our example shows a postcrash equilibrium price of $p_0 = 0.64$: a 30 percent drop from its previous closing price in Figure 3. While the market on October 19 did not fall quite this far, it also is the case that many hedgers scaled back the size of their hedging programs in the face of extraordinarily high transactions costs.²³

A similar story could be told about the 1929 crash. The only difference is that portfolio-insurance hedging would be replaced by stop-loss hedging. While less exact in delivering desired results, stop-loss orders have the effect of increasing liquidity supply as prices fall. It is no accident that investigators focused on the role of stop-loss orders and margined stock buying, since the latter forced additional stop-loss selling as the market descended.

It should be emphasized that a crash in our model is not due to a discontinuous change in the underlying information. Rather, the market reaches a critical point, and a "catastrophe" occurs, both in practice and in theory.²⁴ While hedging strategies are an important part of our explanation of the crash, equally important is the market structure which precludes observing these hedging strategies. Figures 1–4 also plot the excess-demand functions associated with partial or complete observability. These "regular" (i.e., not backward-bending) excess-demand functions eliminate the possibility of crashes in our example. Indeed, if π programs had been fully observable, prices would have fallen a modest 1 percent; and the fall would have been about 1.5 percent if supply-informed investors (only) had observed the extent of π sales.

²³For a description of how hedging programs were modified in the presence of high trading costs, see Leland (1988).

²⁴Hal Varian (1979) discusses catastrophe theory and its relation to economic models. Our discontinuity represents a "cusp catastrophe," as discussed in Section VI.

This is not to say that crashes are impossible with partial observability. If we had assumed a 15-percent use of portfolio insurance ($\omega = 0.15$), the excess-demand curve with partial observation would be backward-bending. A 15-percent use would represent over \$500 billion, or more than five times the total amount estimated for formal programs. Shiller's survey suggested that formal portfolio-insurance programs were "the tip of the iceberg" relative to total hedging, so it is possible that the crash could have occurred in our example even with supply-informed traders aware of hedging supply.

C. After the Fall

A new low-price equilibrium is established after the crash. If information about future prices now reverses itself, returning it to precrash levels of optimism, will the market rebound to its former level? The answer is no. In Figure 4, a small rightward shift of the excess-demand function will lead to a small increase in equilibrium price p_0 from the 0.64 level. Even if the upper branch of the excess-demand curve intersects the zero-excess-demand line, implying the possibility of multiple equilibria, the lower equilibrium price is locally stable and can be expected to prevail. Since the slope of the excess-demand curve is less steep at $p_0 = 0.64$ than just before the crash (when $p_0 = 0.9275$), price volatility will return to lower levels.

Eventually, if information becomes still more favorable (to $p - \bar{p} = 0.026$, well above precrash levels) and if the hedging function π remains the same as before the crash, the excess-demand curve will shift sufficiently to the right such that its lower branch is just tangent to the vertical zero-excess-demand line (see Fig. 5). This will be accompanied by higher volatility. Any further increase in future price expectations could lead to an upward jump in prices: a "meltup" rather than a meltdown. In our example, the discontinuous jump would commence at $p_0 = 0.74$ (15 percent above the market low) and jump to $p_0 = 1.043$.

Perhaps such an upward jump is possible only in the mind of the theorist. However,

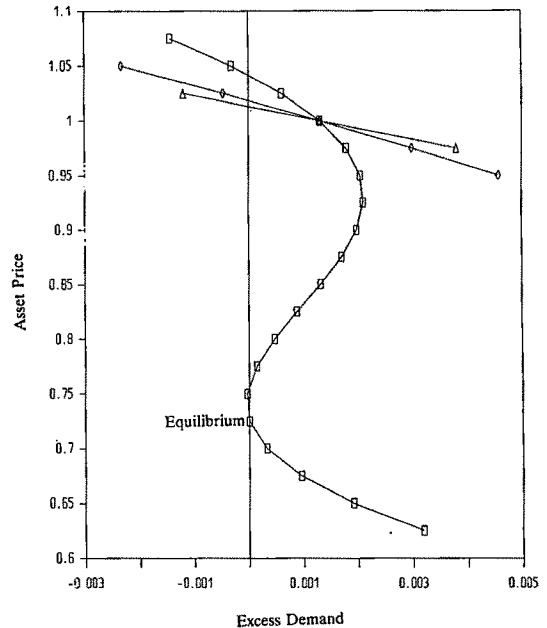


FIGURE 5. AGGREGATE EXCESS DEMAND WITH 2.6-PERCENT UPWARD SHOCK. SYMBOLS: \square = UNOBSERVED; \diamond = PARTIALLY OBSERVED; \triangle = ANTICIPATED

over the period 1928–88, 22 of the 38 one-day stock market moves that exceeded 7 percent were upward jumps, and the financial press occasionally remarks on such a possibility (see Anise Wallace, 1989).

Figure 6 graphs the equilibrium price function relating p_0 to the future price surprise, $p - \bar{p}$, for the three different observability cases. For the case with unobserved π , we see that the point of discontinuity on the upper branch of the function is at $p_0 = 0.927$, and the discontinuity on the lower branch is at $p_0 = 0.740$. For the other two cases, there are no discontinuities given our example's parameters.

Future price surprises are not the only possible sources of discontinuous price behavior. A random liquidity-supply shock could also lead to discontinuous behavior. But whatever the cause, the critical price (i.e., where the discontinuity occurs) will remain the same. This leads us to examine the general nature of critical points: when do they occur, and what determines their level?

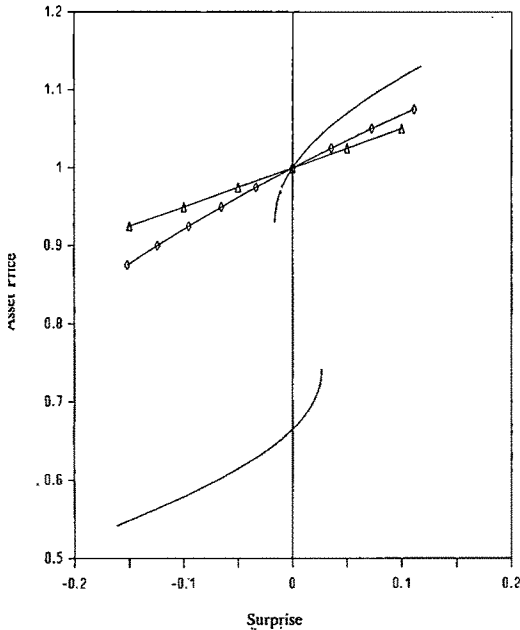


FIGURE 6. EQUILIBRIUM PRICE AS A FUNCTION OF $p - E(p)$. SYMBOLS: — = UNOBSERVED; \diamond = PARTIALLY OBSERVED; Δ = ANTICIPATED

VI. Discontinuities: Some General Results

We now characterize price levels at which the price function becomes discontinuous and the minimum amount of hedging with which a "crash" can occur. These critical points depend upon the extent to which hedging can be observed.

First consider the case in which hedging strategies are unobservable: investors are unaware of hedging strategies and thus do not distinguish them from unobservable liquidity trades. The equilibrium price p_0 is the price level for which excess demand is equal to zero [eq. (5iii)].

Discontinuities will occur if the root (or roots) of (5iii) are discontinuous functions of the variables p , L , and S . Since excess demand is continuous and differentiable in p_0 , discontinuities will take place at points where the function reaches an extremum. Differentiating (5iii) with respect to p_0 yields

$$\frac{\delta XD_U}{\delta p_0} = - \left(\frac{1}{FH} + \pi' \right)$$

where $F = 1 - Z/\Sigma > 0$ is the coefficient that obtains in the case of no hedging (agents are unaware of hedging in this case) and where

$$\pi' = -\omega \frac{N'(d_1)}{p_0 \sigma} (< 0).$$

The derivative of the demand for hedging (π') tends to zero as prices become large and as prices become small; hence, the derivative of excess demand is negative at both very high and very low prices.

The equilibrium price function will be discontinuous if and only if there exists a p_0 such that $(1 + FH\pi') < 0$. This implies prices at which the excess-demand curve is back-wedging and also implies that

$$(6) \quad 1 + FH\pi' = 0$$

admits a solution. Then, since π has a unique inflection point in this case, equation (6) has two solutions, the critical prices c_1 and c_2 ($c_2 > c_1$). Excess demand is an increasing function of equilibrium price p_0 in the interval (c_1, c_2) and decreasing elsewhere. It can also be shown that the first critical point, c_1 , decreases as FH increases and the second, c_2 , increases as FH increases. Equation (6) has two roots if and only if

$$\omega FH > Ke^{-(3/2)\sigma^2} (2\pi\sigma^2)^{1/2} \equiv \phi_{\min}$$

implying

$$\omega_{\min} = (FH)^{-1} \phi_{\min}.$$

The root is unique (and there is no discontinuity) when equality obtains. ω_{\min} represents the largest proportion of hedgers for which a crash does not occur. Note that ω_{\min} also is the upper bound of ω for which the inverse price function $f^{-1}(p_0)$ is monotonically increasing.

The critical prices, c_1 and c_2 , are given by

$$(7) \quad c_1 = Ke^{-2\sigma^2} e^{-[2\sigma^2 \ln(\omega FH / \phi_{\min})]^{1/2}}$$

$$c_2 = Ke^{-2\sigma^2} e^{[2\sigma^2 \ln(\omega FH / \phi_{\min})]^{1/2}}$$

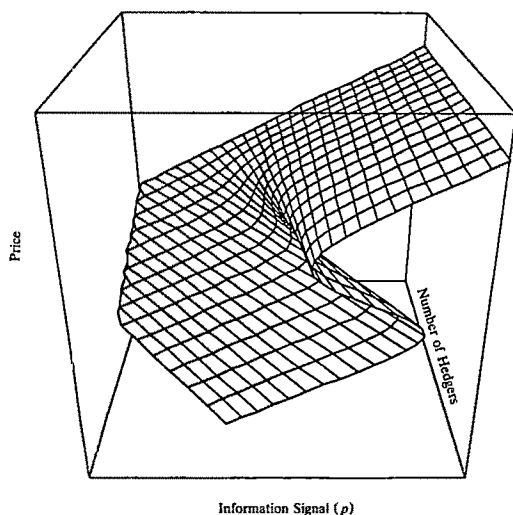


FIGURE 7. EQUILIBRIUM PRICE AS A FUNCTION OF FUNDAMENTALS AND THE NUMBER OF HEDGERS

The difference $c_2 - c_1$ is the range of prices p_0 for which no stable equilibrium exists; however, the amount of the price drop when c_2 is reached from above is larger than this difference. In our base case, the discontinuity occurs for a value of ϕ_{\min} of 0.425, which is reached for $\omega_{\min} = 4.26$ percent. This percentage of hedgers can create a market crash in our example. Conversely, the market meltup takes place when the equilibrium price reaches c_1 from below and the price jumps to a level higher than c_2 (see Fig. 5). Figure 7 graphs equilibrium price as a function of information signals and the fraction of hedgers ω . The graph indicates a "cusp catastrophe": for low values of ω , there is a unique price equilibrium for each signal realization; for higher values, three distinct equilibria exist in the "fold" area. The fold points correspond to the critical points of Figure 6. The cusp point corresponds to ω_{\min} . Note that the interval (c_1, c_2) corresponds to the interval over which the function $p_0 + FH\pi(p_0)$ is decreasing in p_0 and, therefore, is the range of prices over which the inverse price function $f^{-1}(\cdot)$ is decreasing and multiple price equilibria exist.

Now consider the case in which there is partial observation of hedging strategies: supply-informed investors can observe the

sum of S and π . The same reasoning leads to an equation analogous to (6):

$$1 + FI\pi' = 0$$

implying

$$\omega_{\min} = (FI)^{-1} \phi_{\min}.$$

The critical points are obtained by substituting I for H in (7). Since $H > I$ in all cases, the minimum fraction ω_{\min} (10.4 percent in our example) is higher than in the previous case, c_1 is larger, c_2 smaller, and ceteris paribus the price drop is smaller.

Finally, in the fully observed case, identical results obtain provided that FH is replaced with Z : discontinuities require $1 + Z\pi' < 0$. Note that this does not appear to be a major restriction: ω would have to be enormous (over 100 percent).²⁵

In summary, crashes are most likely to occur in the unobserved case, since the inequality is satisfied for the lowest values of ω_{\min} . Because the critical-point difference $c_2 - c_1$ is greatest in this environment, the "crashes" associated with this environment will also be the largest.

VII. Making Markets More Stable

Our analysis suggests that unobserved hedging strategies can destabilize a market, leading to greater volatility and ultimately to a crash. Are there private or governmental policies that would lessen the chances of such an event in the future?

Outlawing hedging strategies is one such possibility, but it is neither practical nor desirable. It is not practical because it is not enforceable. An investor following a stop-loss or portfolio-insurance hedging strategy can always claim he is doing so for other reasons: an anticipated expense, a forecast of weak markets, etc. Short of prohibiting selling for any reason, it is impractical to

²⁵This means that selling by hedgers following a Black-Scholes put-option-replicating strategy would be met by the buying of investors as prices fell continuously, even if hedgers' selling (as prices fell to zero) were 100 percent of initial supply. Indeed, hedge selling would have to be 10 times more intensive than this before price could fall discontinuously.

prohibit selling for hedging purposes. Nor would it be desirable. Investors are willing to participate in a market because they can sell whenever they wish to, including for risk-avoidance purposes.

We note that the market is partially self-correcting. Stop-loss and dynamic hedging strategies are fully effective only when prices move continuously. The possibility of a crash will limit the use of dynamic protection strategies. Of course, if relatively few investors follow such strategies, crashes are unlikely to occur.

Portfolio protection is a legitimate aim of private investors. Is there a way in which investors can achieve protection without contributing to—or suffering from—discontinuous markets? Our analysis provides some clues. The most important result is that widespread knowledge of dynamic hedging usage can minimize its impact on markets. The preceding section showed that the unobserved hedging, which created a 30-percent crash in market prices, would have less than a 1-percent impact on prices if it were observed by all investors. Does this seem preposterous? Some postcrash evidence suggests that it is not. On October 19, 1988, exactly one year after the crash, the Japanese government sold over \$24 billion of a single stock, Nippon Telephone and Telegraph (NT&T). This was four times the amount of all stocks that portfolio insurers had sold the year before. Yet NT&T stock did not decline by a significant amount (either at sale or at the time of initial announcement), because investors had prior knowledge that the sale did not reflect an informational change. Interestingly, portfolio insurers were anxious to disseminate information about their trading requirements prior to the crash, but events happened more quickly than regulatory approval.²⁶

An alternative is for hedgers to use static instruments that provide the same results as dynamic hedging strategies. For example, put options provide protection without requiring further trading. They would seem to be the ideal instrument to avoid the problems of trading in uninformed (and therefore illiquid) markets. A criticism of this argument is that it simply pushes the problem back one level: the sellers of the put option will need to protect themselves through a dynamic hedging strategy. Even if this is true, however, at least there will be publicly available information about the number of outstanding put options. Astute observers can “reverse engineer” the dynamic strategies that the open interest in such options imply. If this information is widely disseminated, we will have nearly universal observation of π strategies.

Short of all investors being aware of hedging plans, our analysis also shows that the stability of markets is strongly affected by the number of supply-informed traders who can observe these plans. These market-makers play a role far beyond their numbers in increasing market liquidity. The crash that occurred in our example with no investors observing hedging could have been prevented if there had been as few as 0.03 percent supply-informed investors (given $\omega = 0.05$) observing hedging supply π .

To the extent that stock-exchange specialists have privileged access to information on the nature of order flows, they play a key role in providing stability. Rules that limit free entry to this activity will leave markets considerably more vulnerable than otherwise. Electronic “open books” should be a seriously considered reform, and other forms of market organization (such as single-price auctions) should be examined.

Low margin requirements in stock or derivatives markets can lead to an increased level of forced margin sales as prices fall. In effect, low margins increase the likely amount of stop-loss sales. If the extent of forced margin sales is difficult to observe, low margin requirements could increase the market’s vulnerability to crashes.

Would price limits help? The answer is no—unless such limits (and the trading halts caused by their being reached) permitted

²⁶ With the assistance of a major portfolio-insurance firm (LOR), the New York Futures Exchange (NYFE) had requested the right to publicize large futures sales in advance. The theory behind the request was that preannouncement would allow time for the market to organize a competitive response. Prior to the crash, the NYFE proposal had been withdrawn, reportedly because there were insufficient means of electronically disseminating the information.

better dissemination of information on hedgers' selling. Absent this, price limits would only delay the ultimate crash by a bit, without modifying its magnitude. Certainly the market did not seem to benefit from the "trading halt" created by the weekend of October 17-18.

VIII. Conclusion

We have shown that information differences among market participants can cause financial markets to be relatively illiquid. A small unobserved supply shock can create a large fall in prices. This is because the fall in prices affects investors' expectations as well as their budgets. Traditional models which do not recognize that many investors are poorly informed will grossly overestimate the liquidity of stock markets.

A consequence of diminished liquidity is that even relatively small unobserved trades by hedging programs can have a destabilizing effect. We developed an example in which a market crash occurred when only 5 percent of investors were following a hedging program replicating a put option.

Our model suggests how a crash caused by hedging in this country could be propagated to foreign markets, even when these markets do not have hedging programs such as portfolio insurance. Foreign investors, observing the large price drop in the U.S. market but ignorant of the extent of hedging in that market, rationally infer that significant negative information must have been received by U.S. investors. To the extent that this information is also significant for their own markets, foreign investors revise downward their expectations, causing prices to fall globally.

Our model also indicates policies to minimize the chance of future crashes. These include the wide dissemination of knowledge about hedgers' actions, marginal positions, and the use of put options or related securities that provide hedging without requiring dynamic trading. This recommendation supports a similar contention by Grossman (1988a). Allowing wider access to the information in specialists' books might also help to stabilize the market. In contrast, price limits are unlikely to have useful ef-

fects unless they are combined with greater dissemination of trading information at the time limits are reached.

APPENDIX

Notation (with Example Parameters in Parentheses)

Prices

p_0 :	current equilibrium price
p :	realized end-of-period price
\bar{p} :	unconditional expected end-of-period price (1.06)
\bar{p}_i :	investor i 's conditional expectation of end-of-period price
Σ :	unconditional variance of end-of-period price (0.08)
Z_j :	class j investor-conditional variance of p
Z :	market power-weighted average conditional variance of p

Information

m :	supply of shares divided by the sum of risk-tolerance coefficients; expectation \bar{m} (1.503), variance Σ_m (0.00034)
p'_i :	$p + \varepsilon_i$ price signal observed by investor i in class I
ε_i :	price signal noise, uncorrelated across investors, uncorrelated with other random variables; <i>ex ante</i> variance Σ_ε (0.4)
S :	liquidity supply observed by investors SI; mean 0 and variance Σ_S (0.00017)
L :	unobserved liquidity supply; mean 0 and variance Σ_L (0.00017); L and S are independent

Investors

SI:	supply-informed investor class; observe p_0 and S
I:	price-informed investor class; observe p_0 and p'_i
U:	uninformed investor class; observe p_0
j :	investor class SI, I, or U
a_j :	investor-class j risk tolerance
w_j :	number of investors in class j

- k_j : relative market power of class j ;
 ratio of the products of w_j and a_j to the sum across classes:
 $k_j \equiv a_j w_j / \sum a_j w_j$ ($k_I = 0.02$,
 $k_{SI} = 0.005$, $k_U = 0.975$)
 $\pi(p_0)$: hedging share supply
 ω : fraction of share total hedged
 (5 percent)

PROOF OF THEOREM 1:

We will assume that investors believe the function f^{-1} to be well-defined (i.e., a given equilibrium price level p_0 obtains for only one possible realization of the argument of the function f). Subsequently we will show that this belief is confirmed in equilibrium. The variance-covariance matrix V of the three-signal vector

$$\begin{bmatrix} p' \\ S \\ f^{-1}(p_0) \end{bmatrix}$$

and the covariance vector W of the signal vector with the future price are given by

$$V = \begin{bmatrix} \Sigma + \Sigma_\epsilon & 0 & \Sigma \\ 0 & \Sigma_S & -I\Sigma_S \\ \Sigma & -I\Sigma_S & \Sigma + H^2\Sigma_L + I^2\Sigma_S \end{bmatrix}$$

$$W = \begin{bmatrix} \Sigma \\ 0 \\ \Sigma \end{bmatrix}$$

For simplicity, we have omitted the subscript i (for investor i) of p' . The distribution of end-of-period prices conditional on all three signals is normal with expectation \bar{p}_Π and variance Z_Π . Defining $[A_\Pi, B_\Pi, C_\Pi] \equiv W^\top V^{-1}$, where W^\top denotes the transpose of W , leads to

$$\begin{aligned} \bar{p}_\Pi &= \bar{p} + W^\top V^{-1} \begin{bmatrix} p' - \bar{p} \\ S \\ f^{-1}(p_0) \end{bmatrix} \\ &= \bar{p} + [A_\Pi, B_\Pi, C_\Pi] \begin{bmatrix} p' - \bar{p} \\ S \\ f^{-1}(p_0) \end{bmatrix} \\ Z_\Pi &= \Sigma - \text{Cov}\{p, [p', S, f^{-1}(p_0)]\}^\top \\ &\quad V^{-1} \text{Cov}\{p, [p', S, f^{-1}(p_0)]\} \\ Z_\Pi &= \Sigma - (A_\Pi \Sigma + C_\Pi \Sigma) \end{aligned}$$

(see, e.g., Morris DeGroot, 1975). Straightforward and lengthy manipulation of the equations leads to

$$\begin{aligned} Z_\Pi &= \left[\frac{1}{\Sigma} + \frac{1}{\Sigma_\epsilon} + \frac{1}{H^2\Sigma_L} \right]^{-1} \\ Z_\Pi^{-1}A_\Pi &= \frac{1}{\Sigma_\epsilon} \\ Z_\Pi^{-1}B_\Pi &= \frac{I}{H^2\Sigma_L} \\ Z_\Pi^{-1}C_\Pi &= \frac{1}{H^2\Sigma_L} \end{aligned}$$

These parameters would obtain for an investor who could observe all the signals. To derive the corresponding parameters for the supply-informed investors (SI) it suffices to take the limit of Σ_ϵ at infinity. For investors I and U, who do not observe the signal S , the parameters are obtained by replacing $H^2\Sigma_L$, the contribution to the variance of $f^{-1}(\cdot)$ due to unobserved liquidity trading, with $H^2\Sigma_L + I^2\Sigma_S$ in the expression for the corresponding parameters for Π and SI, respectively. This yields

$$\begin{aligned} Z_{SI} &= \left(\frac{1}{\Sigma} + \frac{1}{H^2\Sigma_L} \right)^{-1} \\ Z_{SI}^{-1}A_{SI} &= 0 \\ Z_{SI}^{-1}B_{SI} &= \frac{I}{H^2\Sigma_L} \\ Z_{SI}^{-1}C_{SI} &= \frac{1}{H^2\Sigma_L} \\ Z_I &= \left(\frac{1}{\Sigma} + \frac{1}{\Sigma_\epsilon} + \frac{1}{H^2\Sigma_L + I^2\Sigma_S} \right)^{-1} \\ Z_I^{-1}A_I &= \frac{1}{\Sigma_\epsilon} \\ Z_I^{-1}B_I &= 0 \\ Z_I^{-1}C_I &= \frac{1}{H^2\Sigma_L + I^2\Sigma_S} \\ Z_U &= \left(\frac{1}{\Sigma} + \frac{1}{H^2\Sigma_L + I^2\Sigma_S} \right)^{-1} \\ Z_U^{-1}A_U &= 0 \\ Z_U^{-1}B_U &= 0 \end{aligned}$$

$$Z_U^{-1}C_U = \frac{1}{H^2\Sigma_L + I^2\Sigma_S}.$$

The corresponding market power weighted averages are given by

$$Z^{-1} \equiv \sum_j k_j Z_j^{-1} = \frac{1}{\Sigma} + \frac{k_1}{\Sigma_\epsilon} + \frac{H^2\Sigma_L + k_{SI}I^2\Sigma_S}{H^2\Sigma_L(H^2\Sigma_L + I^2\Sigma_S)}$$

$$A \equiv \sum_j k_j Z_j^{-1}A_j = \frac{k_1}{\Sigma_\epsilon}$$

$$B \equiv \sum_j k_j Z_j^{-1}B_j = k_{SI} \frac{I}{H^2\Sigma_L}$$

$$C \equiv \sum_j k_j Z_j^{-1}C_j \\ = k_{SI} \frac{1}{H^2\Sigma_L} + \frac{k_1 + k_U}{H^2\Sigma_L + I^2\Sigma_S}.$$

The total demand for shares of the three classes of investors is equal to the total supply plus hedging supply:

$$(A1) \quad \sum_j k_j Z_j^{-1}(\bar{p}_j - p_0) = m + \pi.$$

Reorganizing terms yields, at the limit of economies with an infinite number of agents,

$$(A2) \quad \frac{Z^{-1}p_0 + \pi - C f^{-1}(p_0) - Z^{-1}\bar{p} + \bar{m}}{A} \\ = p - \bar{p} - \frac{1}{A}L - \frac{1-B}{A}S.$$

This equation is consistent with equation (3) if and only if the following set of equations holds:

$$H = \frac{1}{A} \quad I = \frac{1-B}{A}$$

$$f^{-1}(p_0) = \frac{Z^{-1}p_0 + \pi - Z^{-1}\bar{p} + \bar{m}}{A + C}$$

Substituting A , B , and C yields the unique

solution for H , I , and f^{-1} :

$$(A3) \quad H = \frac{\Sigma_\epsilon}{k_1} \quad I = H - \frac{Hk_{SI}}{H\Sigma_L + k_{SI}}$$

$$f^{-1}(p_0) = \frac{Z^{-1}p_0 + \pi - Z^{-1}\bar{p} + \bar{m}}{Z^{-1} - \Sigma^{-1}}.$$

The solution f^{-1} is a well-defined function, as asserted above.

PROOF OF PROPOSITIONS 1 AND 2:

The function f^{-1} is continuous. Consequently, the function f is well-defined and continuous if and only if f^{-1} , or equivalently $Z^{-1}p_0 + \pi$, is strictly monotonic. If $\pi(p_0) = 0$, f^{-1} is strictly monotonic, since $Z^{-1} > 0$; hence f is well-defined and continuous.

Excess Demand. Substitution of the solutions in equation (A2) yields the excess demand (demand minus supply):

$$XD_A = \frac{1}{H} \left[p - \bar{p} - HL - IS \right. \\ \left. + \frac{Z^{-1}(\bar{p} - p_0) - (\bar{m} + \pi)}{Z^{-1} - \Sigma^{-1}} \right].$$

The Linear Case. When the demand stemming from dynamic strategies is linear in p_0 (i.e., π' is constant), $f(\cdot)$ is a linear function. In the case of no hedging supply ($\pi = 0$), the equilibrium price p_0 is given by

$$(A4) \quad p_0 = \frac{Z^{-1} - \Sigma^{-1}}{Z^{-1}} (p - \bar{p} - HL - IS) \\ + \bar{p} - Z\bar{m}.$$

We will denote by F the slope of the function f ; in this case, $F = 1 - Z/\Sigma$. In the context of our example, we have $Z^{-1} = 25.06$ and

$$p_0 = 0.5 (p - 1.06 - 19.95L - 8.14S) + 1.$$

9,6015

PROOF OF PROPOSITION 3:

We first restrict our attention to the domain of p_0 where $f^{-1}(p_0)$ is strictly increasing. From (A3), $f^{-1}(p_0)$ is also differentiable, with derivative $Z^{-1} + \pi'(p_0) > 0$ over this domain. As hedging activity $\pi(p_0)$ increases, the derivative of $f^{-1}(p_0)$ decreases, since $\pi' < 0$. Therefore, the derivative of $f(\cdot)$ becomes larger, and the current price becomes more sensitive to changes in the signals. Since the signal volatility is exogenous, this in turn implies that the current price is more volatile. For sufficiently large hedging activity, f^{-1} actually decreases over the range of prices p_0 for which

$$(A5) \quad -\pi'(p_0) > Z^{-1}.$$

Therefore, $f(\cdot)$ is multivalued, and discontinuities within the set of stable equilibria can occur, as demonstrated in the example of Section VI.

PROOF OF THEOREM 2:

The proof closely follows that of Theorem 1. The difference consists in the agents' different beliefs about the structure of equilibrium prices. In the first case, supply-informed agents (SI) are aware of hedging strategies and of their impact on prices. SI agents know f^{-1} , the actual inverse price function which obtains in equilibrium. Other agents, ignorant of the presence of hedgers, think that the linear functional form holds. We assume that SI agents believe the coefficients H and I to be unchanged and show that it indeed holds in equilibrium. Equation (A1) still holds, and a similar manipulation leads to the analog of (A2):

$$(A6) \quad \frac{\frac{Z^{-1}(p_0 - (\bar{p} - Z\bar{m}))}{Z^{-1} - \Sigma^{-1}} \left(A + \frac{B}{I} \right) + \pi - C f^{-1}(p_0)}{A + \frac{B}{I}} \\ = p - \bar{p} - \frac{1}{A} L - \frac{1-B}{A} S.$$

Hence, the parameters H and I are unchanged, and $f^{-1}(p_0)$ is given by the left-hand side of equation (A6), because SI agents know the true inverse equilibrium

price function $f^{-1}(\cdot)$. The solution $f^{-1}(p_0)$ is given by

$$(A7) \quad f^{-1}(p_0) \\ = \frac{Z^{-1}(p_0 - (\bar{p} - Z\bar{m}))}{Z^{-1} - \Sigma^{-1}} + I\pi.$$

If hedging activity is totally unobserved, similar derivations yield the same parameters H and I as before, and the new inverse equilibrium price function

$$(A8) \quad f^{-1}(p_0) \\ = \frac{Z^{-1}(p_0 - (\bar{p} - Z\bar{m}))}{Z^{-1} - \Sigma^{-1}} + H\pi.$$

PROOF OF PROPOSITIONS 4 AND 5:

The derivative of the inverse equilibrium price function $d(f^{-1})/dp_0$ is equal to $F^{-1} + (Z^{-1} - \Sigma^{-1})^{-1}\pi'$ in the fully observed case, to $F^{-1} + I\pi'$ in the partially observed case, and to $F^{-1} + H\pi'$ in the unobserved case [eqs. (A3), (A7), and (A8)]. It is smallest for the unobserved hedging activity case, and it is smaller under partial observation than in the fully observed case, because π' is negative and $H > I > (Z^{-1} - \Sigma^{-1})^{-1}$ [by combining the definition of Z^{-1} and eq. (A3)]. This implies that the derivative of f (and therefore the volatility of p_0) is largest in the case of unobserved hedging activity and least in the case when hedging is fully observed. The derivatives are negative if $-\pi' > Z^{-1} = (Z^{-1} - \Sigma^{-1})F^{-1}$ in the observed case, if $-\pi' > (FI)^{-1}$ in the partially observed case, and if $-\pi' > (FH)^{-1}$ in the unobserved case. Hence, as hedging activity increases, discontinuities appear first in the unobserved case, then in the partially observed case, and finally in the perfectly observed case.

REFERENCES

- Admati, Anat R., "A Noisy Rational Expectations Equilibrium for Multi-Asset Securities Markets," *Econometrica*, May 1985, 53, 629-57.

- Black, Fischer**, "An Equilibrium Model of the Crash," in *NBER Macroeconomics Annual 1988*, Cambridge, MA: MIT Press, 1988, 269-75.
- _____, and **Scholes, Myron S.**, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy*, May-June 1973, 81, 637-54.
- Brady, Nicholas et al.**, *Report of the Presidential Task Force on Market Mechanisms*, Washington: U.S. Government Printing Office, 1988.
- Brennan, Michael J. and Schwartz, Eduardo S.**, "Portfolio Insurance and Financial Market Equilibrium," *Journal of Business*, October 1989, 62, 455-76.
- Cutler, D., Poterba, James M. and Summers, Lawrence H.**, "What Moves Stock Prices?" *Journal of Portfolio Management*, Spring 1989, 15, 4-12.
- DeGroot, Morris H.**, *Probability and Statistics*, Reading, MA: Addison-Wesley, 1975.
- Diamond, Douglas W. and Verrecchia, Robert E.**, "Information Aggregation in a Noisy Rational Expectations Economy," *Journal of Financial Economics*, Summer 1981, 9, 221-35.
- Gennotte, Gerard**, "A Rational Expectations Asset Pricing Model," Ph.D. Dissertation, Sloan School of Management, Massachusetts Institute of Technology, 1985.
- _____, and **Leland, Hayne E.**, "On the Stock Market Crash and Portfolio Insurance," paper delivered to the American Finance Association, December 1987.
- Glosten, Lawrence R. and Milgrom, Paul R.**, "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics*, March 1985, 14, 71-100.
- Grossman, Sanford J.**, "On the Efficiency of Competitive Stock Markets Where Agents Have Diverse Information," *Journal of Finance*, May 1976, 31, 573-85.
- _____, (1988a) "An Analysis of the Implications for Stock and Futures Price Volatility of Program Trading and Dynamic Hedging Strategies," *Journal of Business*, July 1988, 61, 275-98.
- _____, (1988b) "Insurance Seen and Unseen: The Impact on Markets," *Journal of Portfolio Management*, Summer 1988, 14, 5-8.
- _____, and **Stiglitz, Joseph E.**, "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, June 1980, 70, 393-408.
- Hellwig, Martin**, "On the Aggregation of Information in Competitive Markets," *Journal of Economic Theory*, June 1980, 22, 477-98.
- Henriksson, Roy D.**, "Market Timing and Mutual Fund Performance: An Empirical Investigation," *Journal of Business*, July 1984, 57, 73-96.
- Ibbotson Associates, Stocks, Bonds, Bills, and Inflation: 1985 Yearbook**, Chicago: Ibbotson Associates, 1985.
- Kyle, Albert S.**, "Continuous Auctions and Insider Trading," *Econometrica*, November 1985, 53, 1315-35.
- Leland, Hayne E.**, "Portfolio Insurance and October 19th," *California Management Review*, Summer 1988, 30, 80-9.
- _____, and **Rubinstein, Mark**, "Comments on the Market Crash: Six Months After," *Journal of Economic Perspectives*, Summer 1988, 2, 45-50.
- Roll, Richard**, "The International Crash of October 1987," in Robert W. Kamphuis, Roger C. Kormendi, and J. W. Henry Watson, eds., *Black Monday and the Future of Financial Markets*, Homewood, IL: Irwin, 1989, 35-70.
- Rubinstein, Mark and Leland, Hayne E.**, "Replicating Options with Positions in Stock and Cash," *Financial Analysts Journal*, July-August 1981, 63-72.
- Shiller, Robert J.**, "Investor Behavior in the October 1987 Stock Market Crash: Survey Evidence," Working Paper, Yale University, November 1987.
- Varian, Hal R.**, "Catastrophe Theory and the Business Cycle," *Economic Inquiry*, January 1979, 17, 14-28.
- Wallace, Anise**, "Nervous Wall Street Fears a 'Melt-Up,'" *New York Times*, May 11, 1989, p. C8.

Insider Trading In A Rational Expectations Economy

By LAWRENCE M. AUSUBEL*

It is often argued that efficiency considerations require society to freely permit insider trading. In this article, an opposing efficiency argument is formalized. The model incorporates an investment stage followed by a trading stage. If "outsiders" expect "insiders" to take advantage of them in trading, outsiders will reduce their investment. The insiders' loss from this diminished investor confidence may more than offset their trading gains. Consequently, a prohibition on insider trading may effect a Pareto improvement. Insiders are made better off if they can precommit not to trade on their privileged information; government regulation accomplishes exactly this. (JEL 022, 026, 313)

The traditional rationale articulated for insider trading regulation and other securities law is that such rules promote confidence in markets. Indeed, President Franklin D. Roosevelt justified the first major U.S. securities legislation by saying: "It should give impetus to honest dealing in securities and thereby bring back public confidence".¹ Similar language is still invoked half a century later, in connection with enforcement efforts against insider trading and in proposals for tightened stock market regulation.

*Department of Managerial Economics and Decision Sciences, J. L. Kellogg Graduate School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208. My research received the gracious support of the Lynde and Harry Bradley Foundation, the Kellogg School's Banking Research Center, and National Science Foundation Grant SES-86-19012. I thank Laurie Bagwell, Mike Fishman, Julie Nelson, Matt Spiegel, and three anonymous referees, as well as seminar participants at the 1988 Winter Meetings of the Econometric Society, the Midwest Mathematical Economics Meetings (April 1989), Northwestern University, and Indiana University, for helpful comments.

¹77 Congressional Record 937 (March 29, 1933). The quote is taken from President Roosevelt's message to Congress in proposing legislation that became the Securities Act of 1933. This act requires the disclosure of information in connection with the initial offering of a security. The legislative underpinnings of federal insider trading regulation are contained in the closely related Securities Exchange Act of 1934 (see Section I).

Yet the weight of academic law-and-economics commentary has been opposed to the regulation of insider trading. Scholars have argued that permitting trade on the basis of inside information creates desirable incentives and, for a variety of reasons, improves economic efficiency. At the same time, it has been maintained that the fact of prices reflecting information would prevent insiders from earning significant trading profits at the expense of outsiders or that, in any event, outsiders are not harmed by insider trading.

My objective in the current article will be to reformulate the confidence rationale as an economic argument for insider trading regulation. I develop a two-stage model, consisting of an investment stage followed by a trading stage. In the initial period, agents make their investment decisions based on their expected second-period returns, which in turn hinge on whether they will be "insiders" or "outsiders" and on whether insiders will be permitted to trade on their private information. The second period is a pure exchange economy (of endowments determined by the first-period investments) in which it is feasible for insiders to exploit their private information in a partially revealing rational expectations equilibrium.

For many plausible specifications of the model, the outcome when society regulates insider trading is a Pareto improvement over the outcome when insider trading is permit-

ted.² Under such scenarios, economic efficiency would require the banning of insider trading. The intuition for this conclusion, which contradicts most previous economic analyses of insider trading, is as follows. Abolition of insider trading in an exchange situation will typically improve the expected return on investment of outsiders. If the quantity of investment increases in the expected return,³ then insider trading regulation promotes investment by outsiders. To the extent that insiders are helped by increased outside investment,⁴ insiders thus also benefit from insider trading regulation. In other words, insiders are made better off if they can somehow precommit not to trade on their privileged information; government regulation and enforcement accomplish exactly this.

My analysis thus provides an economic formalization of the notion of confidence in markets. Let "confidence" be interpreted as the rational belief by outsiders that their return on investment is not being diluted by insiders' trading. Then, perhaps, the goal of insider trading regulation and securities law truly is to foster confidence in markets. When confidence is promoted, outsiders and insiders may benefit alike.

The article is organized as follows. In Section I, I define insider trading and critically discuss the related literature. Section II provides an overview of the structure and

essential ingredients of the model. In Section III, I formally develop the trading stage when insider trading is permitted; in Section IV, I formally develop the investment stage. Section V modifies the model to treat a regulatory regime in which insider trading is banned. Section VI contains a welfare analysis of insider trading regulation for a set of parameter values in the model economy. Section VII provides my conclusions.

I. A Brief Review

A. Insider Trading Defined

Insider trading occurs when an individual (commonly called an "insider") buys or sells securities on the basis of material, nonpublic information. American law imposes an *abstain or disclose* requirement on insiders. Suppose that an individual has privileged access to corporate information which is not generally available and which materially affects investment decisions concerning the company's stock. Under the doctrine of *In re Cady, Roberts & Co.*⁵ and *SEC v. Texas Gulf Sulphur Co.*,⁶ the insider is required to choose between two options: he may either abstain from engaging in any trading activity in the security in question until such time that the information becomes public; or he may, himself, publicly disclose the information to the marketplace before trading. Failure to abstain or disclose may subject the insider to civil liability and criminal prosecution under Rules 10b-5 or 14e-3, which were promulgated by the Securities and Exchange Commission under rule-making authority granted by Congress in Sections 10(b) and 14(e) of the Securities Exchange Act of 1934.⁷

²The reader should be alerted to the words "many plausible specifications of the model" in this sentence. This article demonstrates that, in *some* specifications, a ban on insider trading effects a Pareto improvement. In others, regulation works to help outsiders but harm insiders. See Sections VI and VII.

³Contemporary government policies designed to promote investment and savings seem to be premised on the notion that the quantity of investment increases in the expected return (i.e., that investment is not a Giffen good).

⁴Traditional corporate insiders (e.g., officers and directors) benefit from outside investment, because this investment is a source of needed capital for their organizations. Nontraditional insiders (e.g., investment bankers and arbitrageurs) also benefit from outside investment, because this investment is the origin of initial public offerings and secondary trades, which again contribute to the insiders' livelihoods.

⁵40 SEC 907 (1961).

⁶401 F.2d 833 (2d Cir. 1968) (en banc), *cert. denied*, 394 U.S. 976 (1969).

⁷Rule 10b-5 is a general prohibition on fraudulent acts and practices connected with the trading of securities: the subsequent case law (beginning with the *Cady, Roberts* and *Texas Gulf Sulphur* cases) has interpreted this rule to proscribe insider trading. Rule 14e-3 is a ban on fraudulent or deceptive acts and practices specifically connected with tender offers.

The term "insider," as used here, refers not merely to a traditional corporate insider but more broadly to any individual whose actions are confined by the insider trading laws. Whether an individual is considered to be an insider, and thus whether he is bound by the abstain-or-disclose requirement, may depend on which of Rules 10b-5 or 14e-3 is being applied. In order to violate Rule 10b-5, the individual must be linked with the firm whose security is traded or with the nonpublic information in such a way that the use of the information in his trading is deemed to breach some fiduciary duty. Under limitations set forth in *Chiarella v. United States*⁸ and affirmed in *Dirks v. SEC*,⁹ the insider's fiduciary duty may derive from: (a) working inside the firm (e.g., employment as an officer or director of the company whose securities are traded); (b) working outside the firm in a capacity which nevertheless leads to an obligation to shareholders (e.g., employment as an investment banker, lawyer, or accountant for the company whose securities are traded); or (c) receiving information from another individual whose conveyance of the information itself constitutes a breach of duty (e.g., receiving a tip from a corporate officer who expects to benefit from the disclosure). Alternatively, under the so-called *misappropriation theory*, if an individual trades on the basis of information misappropriated from its source (typically, taken from the individual's employer), the misuse of information may itself constitute the breach of fiduciary duty that is required for conviction under Rule 10b-5.¹⁰

⁸445 U.S. 222 (1980).

⁹463 U.S. 646 (1983).

¹⁰The misappropriation theory was adopted by the U.S. Court of Appeals, 2nd Circuit, in *United States v. Newman*, 664 F.2d 12 (2d Cir. 1981), *cert. denied*, 104 S.Ct. 193 (1984). It was considered inconclusively by the U.S. Supreme Court in *Carpenter v. United States*, 108 S.Ct. 316 (1987). Divided in a 4-4 vote, the Court failed to overturn the insider trading convictions of *Wall Street Journal* reporter R. Foster Winans (and others) for prepublication trading on the basis of information that would appear in his "Heard on the Street" column.

In contrast, Rule 14e-3 does not contain any duty requirement in its notion of who is an "insider." (It does, however, retain the notion from Rule 10b-5 that "willful misconduct" is a prerequisite to any violation.¹¹) If *any* individual (not necessarily a person who has breached a fiduciary duty) trades while in possession of material, nonpublic information *connected with a tender offer* by another party,¹² he may be subject to prosecution for insider trading under Rule 14e-3.^{13,14}

B. The Classic Law-and-Economics View

The classic law-and-economics view on insider trading can be briefly summarized as follows. Insider trading is banned today out of considerations of *fairness*. In an unregulated environment, insiders might be able to earn trading profits by utilizing information that outsiders cannot legally obtain. Out of some sentimental attachment to fairness, we enact insider trading regulations in order to level the securities market playing field, so that all traders have relatively equal access to information.

Unfortunately, considerations of economic *efficiency* work in the opposite direc-

¹¹*United States v. Chestman*, 704 F. Supp. 451 (S.D.N.Y. 1989); *United States v. Marcus Schloss & Co., Inc.*, 710 F. Supp. 944 (S.D.N.Y. 1989).

¹²A party planning to make a tender offer (or his agent) is permitted to purchase shares on the open market in advance of a public announcement—provided he does not run astray of other provisions of the Williams Act.

¹³It should be added that there exists another federal rule under which (only civil) insider trading liability is possible. A traditional insider (an officer, director, or major shareholder) is liable to his company for any profits he earned from *matched purchases and sales of securities within the same six-month period* (irrespective of whether it can be shown that he possessed material, nonpublic information), under Section 16 of the Securities Exchange Act of 1934. The presumption behind this provision on "short swing" trading seems to be that, whenever an insider buys and sells in close proximity, it is likely to be on the basis of (possibly unidentifiable) private information.

¹⁴The discussion of fiduciary duty contained in the second and third paragraphs of this section is largely drawn from Chapters 3, 6, and 7 of Donald Langevoort (1990).

tion as those of fairness. First, if insiders are permitted to trade freely on their private information, then information becomes more rapidly reflected in securities prices. Insider trading thus contributes to efficient markets and so to allocational efficiency, as proper capital-asset pricing leads to the optimal allocation of capital resources. Second, "profits from insider trading constitute the only effective compensation scheme for entrepreneurial services in large corporations" (Henry Manne, 1966b p. 116). As Manne viewed the world, individuals do little innovation except when they are afforded the opportunity to share in the value they create; in large organizations, insider trading is basically the only mechanism for employees to obtain compensation for their innovations.

Furthermore, the fairness considerations are misplaced, as insider trading is effectively a victimless crime:

The insiders' gain is not made at the expense of anyone. The occasionally voiced objection to insider trading—that someone must be losing the specific money the insiders make—is not true in any relevant sense.

[Manne, 1966a p. 61]

Even if the redistributive concerns are real, they are difficult for economists to evaluate (or are irrelevant) because any profit derived from insider trading is an essentially costless transfer payment. Finally, the fact that insider trading by the company's own management is typically not banned by explicit provisions of the corporate charter may be taken as evidence that governmental insider trading regulation does not enhance shareholder value (Dennis Carlton and Daniel Fischel, 1983).

C. *Some Criticisms of the Classic View*

The model described in the following sections of the article does not directly address some of the above arguments. Instead, I

stake out a new argument which cuts in the opposite direction. Thus, before proceeding with the new model, it may prove useful to review and articulate some direct responses to the classic view.

In a very general sense, there exists a fundamental tension in the viewpoint that insider trading promotes economic efficiency. As Manne recognized in the first sentence of his 1966 treatise (but which remains equally true today), "Probably no aspect of modern corporate life has been more roundly condemned than insider trading." It is somewhat awkward to reconcile his view (of insider trading as the guarantor of efficient markets and the protector of entrepreneurship in the modern corporation) with the almost universal opprobrium that society directs toward practitioners of insider trading.

If insider trading is efficient (or even if insiders as a group benefit from the practice), there remains a political economy puzzle as to why insider trading regulations are ever promulgated. As David Haddock and Jonathan Macey (1987 p. 312) observe:

Modern public-choice theory suggests that regulatory actions, including the decisions of the SEC, will divert wealth from relatively diffuse groups toward more coalesced groups whose members have strong individual interests in the regulation's effect. Yet, if one adopts the conventional view that the battle lines of insider trading regulation are drawn between insiders and ordinary shareholders (or the general public), the SEC would seem to be channeling wealth that otherwise would be captured by a group with relatively cohesive interests (insiders) toward those with extremely weak and diffuse interests (ordinary shareholders or the general public).

The fact that, empirically, we witness prescriptions against the practice becomes much less a mystery if insiders have a group interest in precommitting not to trade on

their private information, as I will argue below.¹⁵

More specifically, Manne's incentive argument has been criticized on account that insider trading, as a compensation device, creates a moral hazard problem.¹⁶ An individual who has the abilities both to generate and to trade on inside information is given the perverse incentive to generate "bad" news, which is easier to create than "good" news yet equally profitable to trade on (by selling short, instead of buying long).¹⁷ Meanwhile, a company-granted call option is probably a more finely tuned instrument for giving an employee a stake in the value of the corporation's stock than is legalized insider trading (also avoiding the moral hazard problem). In any case, the incentive argument would not appear to be especially relevant to the recent rage of insider trading, which has mostly involved market professionals (e.g., investment bankers and arbitrageurs), rather than traditional corporate "insiders" engaging in entrepreneurial activities.

Researchers have also challenged the notion that insider trading necessarily increases the rapidity with which information

becomes reflected in stock prices. Victor Brudney (1979 note 43), Frank Easterbrook (1981) and others have noted that the prospect of insider trading may give corporate insiders an incentive to delay the disclosure of information to the marketplace. In a recent paper, Michael Fishman and Kathleen Hagerty (1989) argue that the presence of insider trading may discourage outsiders (e.g., stock analysts) from independently generating information, perhaps leading to less informative securities prices.

Finally, it has been observed that the failure of firms to ban insider trading on their own does not constitute conclusive evidence that public regulation is inefficient. One retort is offered by Richard Posner (1986 p. 393), who notes that "if the probability of detection is so low that heavy penalties—which private companies are not allowed to impose—would be necessary to curtail the practice, it might not pay companies to try to curtail it." Examples of "heavy penalties" available only to public enforcers include prison terms and lifetime debarment from the securities industry. Moreover, if insider trading is forbidden by the government and if such laws are not expected to change, then the presence of trading restrictions in corporate charters would be redundant and unnecessary.

D. Other Related Literature

There exists a fairly extensive empirical and experimental literature on insider trading. James Lorie and Victor Niederhoffer (1968), Jeffrey Jaffe (1974), Nejat Seyhun (1986), and others have examined the profitability of trading rules based on the actual purchases and sales of corporate officers, directors and major stockholders (who are required, by Section 16 of the Securities Exchange Act of 1934, to report their transactions). The studies have found that insiders can, in fact, earn extranormal trading profits. Extensive (unpublished, but widely publicized) experimental work on insider trading was conducted in the mid-1980's by R. Foster Winans, Dennis Levine, Ivan Boesky, Drexel Burnham Lambert Inc., and others. The experimental studies were able

¹⁵As evidence that insiders may actually wish to quash insider trading, see the recent comments of Arthur Levitt, Jr., chairman of the American Stock Exchange: "If the investor thinks he's not getting a fair share, he's not going to invest and that is going to hurt capital formation in the long run" (quoted in *Business Week*, April 29, 1985, p. 79). Also observe that the Insider Trading and Securities Fraud Enforcement Act of 1988, which greatly increased the monetary penalties and jail terms for these crimes, ultimately gained the backing of the Securities Industry Association, a leading industry group (as reported in *The New York Times*, October 23, 1988, pp. 1, 15). Finally, it is interesting that both the 1988 act and the Insider Trading Sanctions Act of 1984 (which stiffened penalties and plugged the options loophole) passed the U.S. Congress without any dissenting votes.

¹⁶For a longer discussion of the moral hazard problem, see Joel Seligman (1985 pp. 1094–6). For a general discussion of the relative merits of compensating managers by allowing them to trade on private information, see Ronald Dye (1984).

¹⁷It is worth observing here that Section 16 of the Securities Exchange Act of 1934 prohibits short selling by officers and directors of shares in their own company.

to replicate the conclusions that had been reached by the earlier empirical articles.

A long line of theoretical articles has addressed the issue of information revelation in an asymmetrically informed market with a large number of traders. In Beth Allen (1981), Douglas Diamond and Robert Verrecchia (1981), Sanford Grossman and Joseph Stiglitz (1980), James Jordan (1983), Roy Radner (1979), and other models, some or all of the informed agents' private information is revealed to uninformed agents via the inversion of the rational expectations equilibrium price function. I have provided a more thorough review of the microeconomic rational expectations equilibrium literature in the introduction of a previous article (Ausubel, 1990).

Some other theoretical articles have examined information revelation in a market in which only a single agent possesses a relevant piece of information. Some of this literature also explicitly discusses insider trading. Douglas Gale and Martin Hellwig (1987), Richard Kihlstrom and Andrew Postlewaite (1983), Albert Kyle (1985), Jean-Jacques Laffont and Eric S. Maskin (1990), and others have studied the strategic revelation of information, which necessarily becomes an issue in this context.¹⁸

II. An Overview of the Model

Consider a two-period model with two goods, two components to the state of the world, and two types of representative agents, who are denoted *insiders* and *outsiders*. In the first period, which occurs before any agent has received private information, insiders individually decide how much labor to invest in producing good x and outsiders individually decide how much la-

bor to invest in producing good y . Between the first and second periods, insiders are privately informed of the state of the world, while outsiders receive no private information. (The state affects agents by entering into their state-dependent utility functions.) In the second period, insiders and outsiders trade in a pure exchange economy, where agents' "endowments" (which are now treated as exogenous) equal their investment decisions of the first period.

Even in a trading process where insiders are permitted to trade freely on the basis of their private information, outsiders learn at least some of the insiders' information, via the rational expectations equilibrium (REE) price function. However, since there are *two* components to information but only *one* relative price to reveal it, the REE is only *partially revealing*. As a consequence, outsiders make nontrivially inferior decisions to those they would make under full information. Finally, after the second period, all information becomes public, and utilities are realized based on the state of the world and agents' holdings of the two goods at the end of the second period. The timing of events is illustrated in Figure 1.

I now introduce a modeling device which is meant to represent insider trading regulation. Any agent who has been designated an insider is given the choice between two alternatives in the trading round: he may either abstain from trade and, hence, consume precisely his endowment brought forward from the investment round; or he may publicly disclose his information to the marketplace before trading. This modeling device is intended to capture the essential empirical details of the first paragraph of Section I, while abstracting away from any of the legal technicalities in the second and third paragraphs of that section. Under a regulatory regime where insider trading is banned in this manner, each of the representative insiders in the model is induced to disclose his information, so that the trading round is transformed into one of complete information. The outcome of the trading round is thus changed, and in anticipation of that change, the outcome of the investment round also changes.

¹⁸ Insider trading research by Roland Benabou and Guy Laroque (1989), Utpal Bhattacharya and Matthew Spiegel (1989), Jurgens Dennert (1989), and Michael Manove (1989) has also come to my attention since the initial preparation of the current article. These papers use a wide diversity of modeling techniques but share with the current article a healthy degree of skepticism toward the efficiency claims of proponents of insider trading.

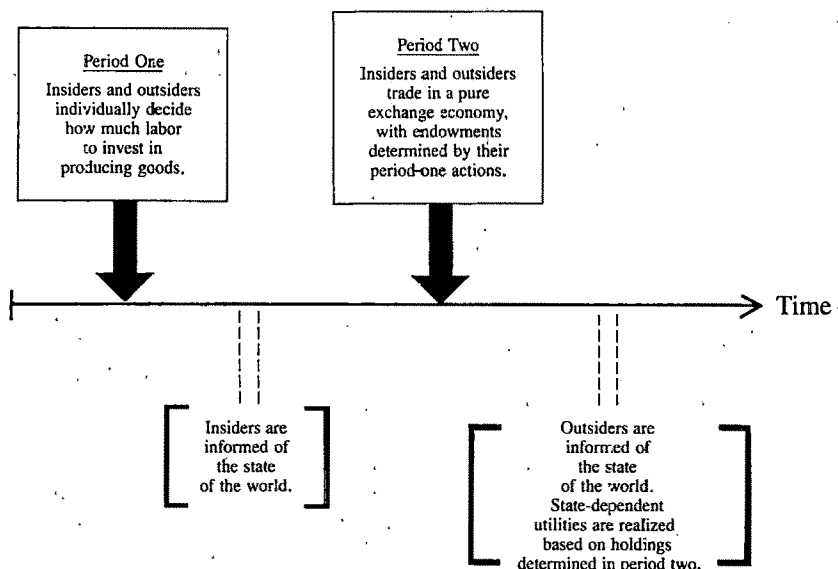


FIGURE 1. TIMING OF EVENTS IN THE INSIDER TRADING MODEL

This analysis makes three basic methodological advances over previous research on insider trading. First, I examine the *ex ante* efficiency of two regulatory regimes, comparing agents' expected utilities before the start of both trading and the underlying investment. Previous analyses have ignored incentive effects on the level of investment activity, only examining utilities preceding a trading round.¹⁹ Second, I fully specify all agents' utility functions and perform a complete welfare comparison. In contrast, earlier work typically specified trading rounds utilizing "noise traders" (who are not explicitly given utility functions) and linear-quadratic functional forms (which yield prices that are sometimes negative); inclusion of either of these features makes welfare analysis problematical. Third, I prove

that my model exhibits a unique equilibrium. Other articles on the subject have frequently chosen a convenient (i.e., linear) solution without resolving whether there exist other equilibria possessing possibly different qualitative characteristics.

Rational expectations equilibrium in a competitive economy specifically models a situation in which there are large numbers of insiders and large numbers of outsiders. In particular, insiders act not only as price-takers but also as information-takers; the modeling technique has insiders ignore that they are affecting the aggregate amount of information available to outsiders when the insiders determine their use of information. This modeling device might fairly well describe a situation such as the November 1988 takeover of Triangle Industries by Pechiney, for which it has been reported that at least eight separate buyers independently bought stock in advance of the acquisition announcement (on the basis of material nonpublic information).²⁰

¹⁹Some earlier commentators have informally taken an *ex ante* view in discussing the fairness of insider trading. For example, Kenneth Scott (1980 p. 809) writes: "The fairness concern proves to have surprisingly little substance when viewed in terms of the game as a whole rather than as a single, isolated play." Just as Scott used an *ex ante* approach to counter the usual fairness argument, I use an *ex ante* approach to counter the usual efficiency argument.

²⁰See, for example, *The New York Times*, January 30, 1989, p. 28. The reader may initially react to the exchange-economy formulation of the trading round in

The REE modeling device would not so well describe a scenario in which a single trader uniquely possessed the private information. However, it seems evident that substitution of a trading round along the latter lines would only exacerbate the efficiency problem. If the insider need not behave as an information-taker, his trading profits would presumably increase, further inhibiting outsiders from investing. Competitive use of information is apparently the most friendly terrain for favoring insider trading; monopolistic use of information would seem only to strengthen the case against it.

It is illuminating to highlight briefly the ingredients of the model that drive the results. First, in order for insiders to profit at the expense of outsiders, it is necessary that the equilibrium of the trading round be only partially revealing. Second, in order for the disparity of information in the trading round to make a difference, it is necessary that insiders and outsiders have different preferences on the underlying commodities.²¹

Section 4 as bearing little relation to a corporate takeover context such as Pechiney/Triangle. The relation becomes much clearer if the terms of the model are reinterpreted. The state-dependent utilities in my model could be viewed as representing the conditional expected utilities that shareholders obtain from holding Triangle's stock. The two components of the state of the world could be thought of as a continuous random variable, β , representing the intrinsic earnings potential of Triangle's capital assets, and a dichotomous random variable, γ , representing whether or not Pechiney is planning a takeover.

²¹This is a natural assumption to make in the market for commodities I use here, since different agents can easily have different preferences over the two commodities. It may not be so obvious to the reader that heterogeneity of preferences is as natural an assumption to make in a market for stocks, since everybody prefers a high return to a low return. In fact, some recent papers have argued that heterogeneity of preferences for stocks is quite natural. Laurie Bagwell (1988) and Yves Balcer and Kenneth Judd (1987) show that, in the presence of a capital gains tax which is imposed upon the realization (rather than the accrual) of a gain, current (taxable) shareholders who purchased the stock at different prices have different objectives. Bagwell and Judd (1988) demonstrate that shareholders with different levels of risk aversion and different marginal propensities to consume have different objectives. Empirical work is also beginning to confirm this assumption. Andrei Schleifer (1986) finds

Third, in order to obtain the strong welfare result that banning insider trading makes *everybody* better off (in some examples), it is useful that: (a) income effects are such that the quantity of investment by outsiders increases as the return on investment increases; and (b) insiders derive some benefit from the investment of outsiders. I will formally outline the model by first giving a description of the second stage of the model and then giving a description of the initial stage.

III. The Trading Stage When Insider Trading Is Permitted

When insider trading is permitted, the second period is modeled as an example of my (1990) partially revealing rational expectations model. The state of the world consists of two independent random variables: a continuous random variable, β , which is uniformly distributed on the unit interval $I \equiv [0, 1]$; and a dichotomous random variable, γ , which takes on the two elements of $\Gamma \equiv \{H, T\}$ ("heads" or "tails") with probabilities h assigned to H and $1 - h$ attached to T. The realization, (β, γ) , is payoff-relevant to agents because it enters into their (state-dependent) utility functions.

Agents are divided into two types, according to their private information. There is a continuum of identical insiders (whose utilities, endowments, and demands are subscripted by 1) and a continuum of identical outsiders (subscripted by 2), each indexed by the unit interval. Insiders privately learn precisely the true realization of (β, γ) between periods one and two, while outsiders do not directly learn the realization until after period two. However, as we shall see, outsiders will indirectly infer some information about the state by observing the price (which, in turn, is influenced by the insiders' actions).

that demand curves for the purchase of stock (which are added to the S&P 500 Index) are downward sloping, as opposed to horizontal. Bagwell (1989) finds that supply curves for the sale of stock (in Dutch-auction repurchases) are upward sloping, as opposed to horizontal.

There are two goods, denoted x and y . Prices for the two goods are assumed to be nonnegative and are normalized to sum to one. I usually only explicitly mention the price of good x , which I denote by the function $p(\cdot, \cdot)$ and the scalar ϕ . A representative insider begins period two with an exogenous endowment (\bar{x}_1, \bar{y}_1) , where $\bar{x}_1, \bar{y}_1 \geq 0$, and trades to a consumption of (x_1, y_1) , where $x_1, y_1 \geq 0$. Similarly, a representative outsider begins period two with an exogenous endowment (\bar{x}_2, \bar{y}_2) and trades to a consumption of (x_2, y_2) . Since each of the two types of agents is indexed by an interval of length one, the aggregate endowments and demands are also given by (\bar{x}_1, \bar{y}_1) , (x_1, y_1) , (\bar{x}_2, \bar{y}_2) , and (x_2, y_2) . In the next section, endowments will be endogenized when we introduce period one; we will then have $\bar{x}_1 > 0$, $\bar{y}_2 > 0$, and $\bar{x}_2 = 0 = \bar{y}_1$.

Agents' utilities derived from consumption are given by state-dependent, Cobb-Douglas utility functions. Let the representative insider's utility function be given by

$$(1) \quad U_1(x_1, y_1; \beta, \gamma) = \begin{cases} x_1^{\alpha_H(\beta)} y_1^{1-\alpha_H(\beta)} & \text{if } \gamma = H \\ x_1^{\alpha_T(\beta)} y_1^{1-\alpha_T(\beta)} & \text{if } \gamma = T \end{cases}$$

where $\alpha_H(\beta) \equiv \beta^{\mu_H}$, $\alpha_T(\beta) \equiv \beta^{\mu_T}$, and μ_H and μ_T are unequal positive constants. Let the representative outsider's utility function be given by

$$(2) \quad U_2(x_2, y_2; \beta, \gamma) = x_2^\beta y_2^{1-\beta}$$

for $\gamma = H, T$.

It can immediately be shown that this competitive model, as specified, does *not* possess any fully revealing rational expectations equilibrium. The reasoning is as follows. Suppose that there were a fully revealing REE, in other words, an equilibrium with the property that an outsider, by observing the market-clearing price, could infer the precise state (β, γ) . Then, the forms of the utility functions in equations (1) and (2) imply that good x is valueless when $\beta = 0$ and that good y is valueless when $\beta = 1$. Consequently, $p(0, H)$ [i.e., the price

when the state is $(0, H)$] would equal zero, $p(1, H)$ would equal one, and the "heads branch" of states would occupy all prices between. Similarly, $p(0, T)$ would equal zero, $p(1, T)$ would equal one, and the "tails branch" of states would occupy all prices between. Now choose any price ϕ such that $0 < \phi < 1$. Then ϕ would be the price associated with two states $[(\beta, H) \text{ and } (\beta', T)]$, contradicting the hypothesis that an outsider could infer the precise state from observing the market-clearing price.

The argument of the previous paragraph further suggests that equilibria of this competitive model will be pairwise revealing (i.e., the outsider should be able to infer from price that the state is one of exactly two possibilities). Indeed, one can prove using a variation of theorem 5 in Ausubel (1990) that any REE of this model is necessarily characterized by a monotone continuous price function²² which, moreover, is pairwise revealing.²³ Taking this fact as given, I will now provide existence and uniqueness results by direct construction.

Observe that the representative insider has full information. For a given price ϕ , let $w_1 \equiv \phi \bar{x}_1 + (1 - \phi) \bar{y}_1$ denote the insider's wealth. Then, in any state (β, γ) , the insider seeks to maximize $U_1(x_1, y_1; \beta, \gamma)$ subject to the budget constraint $\phi x_1 + (1 - \phi) y_1 \leq w_1$. Since utility is Cobb-Douglas, the insider's demand for each good is given by his wealth, divided by the good's price and multiplied by the exponent to which the good's con-

²² It is easy to see that, within the class of monotone and continuous REE price functions, only pairwise-revealing equilibria are possible. Consider any price function, $p(\cdot, \gamma)$ that is monotone and continuous for each of $\gamma = H$ and T . As in the main text, $p(0, H) = p(0, T) = 0$ and $p(1, H) = p(1, T) = 1$; therefore for every $\phi (0 < \phi < 1)$, there exists unique $\beta [0 < \beta < 1]$ and unique $\alpha(\beta) [0 < \alpha(\beta) < 1]$ such that $p(\beta, H) = p(\alpha(\beta), T) = \phi$.

²³ The reasoning behind theorem 5 of Ausubel (1990) establishes that, within the class of (Borel measurable) REE price functions, only pairwise-revealing equilibria are possible. Borel measurability should be considered part of the definition of rational expectations equilibrium. Existence is treated more generally in theorem 2 and corollary 1 of Ausubel (1990).

sumption is raised:

$$(3) \quad x_1(\phi, \bar{x}_1, \bar{y}_1; \beta, \gamma) = [w_1/\phi] \alpha_\gamma(\beta)$$

$$y_1(\phi, \bar{x}_1, \bar{y}_1; \beta, \gamma)$$

$$= [w_1/(1-\phi)] [1 - \alpha_\gamma(\beta)].$$

Define $\alpha(\beta) \equiv \alpha_T^{-1}(\alpha_H(\beta)) = \beta^{\mu_H/\mu_T}$. Using the fact that $\alpha_H(\beta) = \alpha_T(\alpha(\beta))$, it is easy to see that if the price ϕ is identical in states (β, H) and $(\alpha(\beta), T)$, then an insider will display identical demands in those two states.

Now suppose that equilibrium price ϕ is uniquely associated with the states (β, H) and $(\alpha(\beta), T)$. Then, upon observing price ϕ , any outsider cannot determine which of these two states has actually occurred. It is tempting to conclude that the conditional probabilities to place on (β, H) and $(\alpha(\beta), T)$ should merely equal the prior probabilities of H and T , respectively. However, this intuition is misleading and ignores the fact that the two branches of the price function typically have different slopes at a given price observation. Hence, the observation confers additional information. To gain a better intuition for the conditional probabilities, it is helpful to refer to Figure 2. In the Ap-

pendix, it is formally demonstrated that

$$(4) \quad \pi(\beta)$$

$$\equiv \Pr[(\tilde{\beta}, \tilde{\gamma}) = (\beta, H) | p(\tilde{\beta}, \tilde{\gamma}) = p(\beta, H)]$$

$$= \frac{h}{h + [1-h]\alpha'(\beta)}$$

is the correct conditional probability attached to (β, H) , and therefore, $1 - \pi(\beta)$ is the conditional probability attached to the state $(\alpha(\beta), T)$. A representative outsider thus determines his demands $x_2(\phi, \bar{x}_2, \bar{y}_2; \beta, H)$ and $y_2(\phi, \bar{x}_2, \bar{y}_2; \beta, H)$ for the two goods in state (β, H) by solving

$$(5) \quad \max\{\pi(\beta)U_2(x_2, y_2; \beta, H) + [1 - \pi(\beta)]U_2(x_2, y_2; \alpha(\beta), T)\}$$

subject to $\phi x_2 + (1-\phi)y_2 \leq \phi \bar{x}_2 + (1-\phi)\bar{y}_2$

which equally determines his demands $x_2(\phi, \bar{x}_2, \bar{y}_2; \alpha(\beta), T)$ and $y_2(\phi, \bar{x}_2, \bar{y}_2; \alpha(\beta), T)$ in state $(\alpha(\beta), T)$. This maximization problem

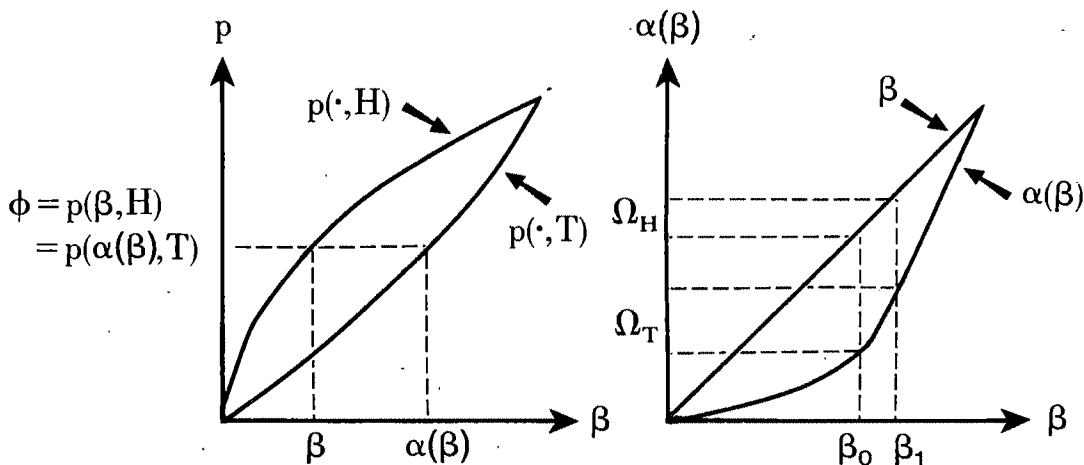


FIGURE 2. INFERENCE FROM A PRICE FUNCTION WITH TWO BRANCHES

yields

$$\begin{aligned}
 (6) \quad x_2(\phi, \bar{x}_2, \bar{y}_2; \beta, H) \\
 &= x_2(\phi, \bar{x}_2, \bar{y}_2; \alpha(\beta), T) \\
 &= [w_2/\phi]\{\pi(\beta)\beta + [1-\pi(\beta)]\alpha(\beta)\} \\
 y_2(\phi, \bar{x}_2, \bar{y}_2; \beta, H) \\
 &= y_2(\phi, \bar{x}_2, \bar{y}_2; \alpha(\beta), T) \\
 &= [w_2/(1-\phi)]\{\pi(\beta)[1-\beta] \\
 &\quad + [1-\pi(\beta)][1-\alpha(\beta)]\}
 \end{aligned}$$

where $w_2 \equiv \phi\bar{x}_2 + (1-\phi)\bar{y}_2$. It may assist the intuition to observe that $\{\pi(\beta)\beta + [1-\pi(\beta)]\alpha(\beta)\}$ and $\{\pi(\beta)[1-\beta] + [1-\pi(\beta)]\alpha(\beta)\}$ are, respectively, the expected values of the exponents to which the consumptions of goods x and y are raised in the outsider's Cobb-Douglas utility function. Hence, similar to (3), an outsider's demand for each good is given by his wealth, divided by the good's price and multiplied by the expected value of the exponent to which the good's consumption is raised.

Suppose that there is a function $p(\beta, \gamma)$ from states of the world to prices that is continuous and strictly monotone in β , with $p(0, H) = p(0, T)$ and $p(1, H) = p(1, T)$. Then for every ϕ satisfying $p(0, H) \leq \phi \leq p(1, H)$, precisely two states are associated with ϕ . We will define a *pairwise revealing rational expectations equilibrium* to be a price function of this form, together with demand functions from equations (3) and (6) that satisfy

$$\begin{aligned}
 (7) \quad x_1(p(\beta, \gamma), \bar{x}_1, \bar{y}_1; \beta, \gamma) \\
 + x_2(p(\beta, \gamma), \bar{x}_2, \bar{y}_2; \beta, \gamma) \equiv \bar{x}_1 + \bar{x}_2
 \end{aligned}$$

for every β ($0 < \beta < 1$) and $\gamma = H, T$; that is to say, agents optimize using the correct rational expectations inference, and markets always clear.²⁴

²⁴Observe that, if the market for good x clears in every state, then by Walras' Law, the market for good y must also clear in every state.

It is straightforward to solve for a closed form of the price function. Substituting (3) and (6) and the definitions of w_1 and w_2 into (7) and solving for $(1-\phi)/\phi$ (i.e., for $[1-p(\beta, H)]/p(\beta, H)$) yields

$$\begin{aligned}
 (8) \quad \psi(\beta) &\equiv \frac{1-p(\beta, H)}{p(\beta, H)} \\
 &= \frac{\bar{x}_1\{1-\alpha_H(\beta)\} + \bar{x}_2\{1-\pi(\beta)\beta - [1-\pi(\beta)]\alpha(\beta)\}}{\bar{y}_1\{\alpha_H(\beta)\} + \bar{y}_2\{\pi(\beta)\beta + [1-\pi(\beta)]\alpha(\beta)\}}
 \end{aligned}$$

where $\pi(\cdot)$ is given by equation (4). It is easy to see that the price function implied by $\psi(\cdot)$ [in eq. (9), below], is always continuous and satisfies $p(0, H) = p(0, T)$ and $p(1, H) = p(1, T)$. If $\psi(\cdot)$ is also a strictly monotone function, then we have constructed a pairwise revealing REE. This establishes the following theorem.

THEOREM 1: *If period one results in endowments $\bar{x}_1, \bar{x}_2, \bar{y}_1$, and \bar{y}_2 such that $\psi(\cdot)$ of equation (8) is strictly monotone in β , then period two has a unique rational expectations equilibrium. In this event, the price function is pairwise revealing and is given by*

$$\begin{aligned}
 (9) \quad p(\beta, \gamma) \\
 = \begin{cases} 1/[1+\psi(\beta)] & \text{if } \gamma = H \\ 1/[1+\psi(\alpha^{-1}(\beta))] & \text{if } \gamma = T \end{cases}
 \end{aligned}$$

where $\alpha^{-1}(\beta) = \beta^{\mu_T/\mu_H}$.

If $\psi(\cdot)$ is not monotone, then period two does not possess any REE.

IV. The Investment Stage When Insider Trading Is Permitted

The first period is modeled quite simply. Agents individually decide how much labor to invest in producing endowment for the second period. At the time of their decisions, agents know whether they will be insiders or outsiders in the second period but do not yet possess any private information (and so they apply the prior distributions on β and $\tilde{\gamma}$). To simplify the subsequent analysis, assume that insiders can only produce good x and outsiders can only pro-

duce good y . The disutility of labor for a representative insider producing x units of endowment is given by $L_1(x)$, and for a representative outsider producing y units of endowment is given by $L_2(y)$. We assign $L_1(\cdot)$ and $L_2(\cdot)$ the following functional forms:

$$(10) \quad L_1(x) = \omega_1 x^{\rho_1} \quad \text{and} \quad L_2(y) = \omega_2 y^{\rho_2}$$

$$\omega_i > 0, \quad \rho_i > 1 \quad (i = 1, 2).$$

The solution concept for the two-stage game will essentially require the play to be a Nash equilibrium in the first period and a rational expectations equilibrium in the second period. Sequential rationality is imposed in the sense that agents in the first period compute payoffs assuming equilibrium play in the second period (i.e., the solution is required to be a backward induction equilibrium).

As indicated above, insiders (and outsiders) are indexed by the unit interval. Hence, if all representative insiders (outsiders) decide in period one to produce \bar{x}_1 (\bar{y}_2), then their aggregate endowment entering period two equals \bar{x}_1 (\bar{y}_2). Moreover, any individual agent's investment decision has absolutely no effect on aggregate endowment (nor on the REE price function), and so he takes aggregate endowments (and the resulting REE price function) as given when selecting his investment.

In order to state the first-period optimization problems, expressions are needed for the second-period payoffs from individual choices of endowment when aggregate endowment is expected to equal (\bar{x}_1, \bar{y}_2) . Let $\bar{V}_1(x)$ ($\bar{V}_2(y)$) denote the *ex ante* expected utility—excluding the $L_i(\cdot)$ term—to a representative insider (outsider) who has carried forward x (y) units of endowment into period two, before learning anything about the state. Formulas for $\bar{V}_1(x)$ and $\bar{V}_2(y)$ are derived in equations (A2)–(A5) of the Appendix.

Now let $X_1(\bar{x}_1, \bar{y}_2)$ signify the optimal endowment for an insider to produce individually in the first period if he expects aggregate endowment in the second period to

equal (\bar{x}_1, \bar{y}_2) . Then $X_1(\bar{x}_1, \bar{y}_2)$ solves

$$(11) \quad \max_{x \geq 0} \{-L_1(x) + \bar{V}_1(x)\}$$

where $\bar{V}_1(\cdot)$ is derived using the price function from aggregate endowments (\bar{x}_1, \bar{y}_2) . Similarly, let $Y_2(\bar{x}_1, \bar{y}_2)$ signify the optimal endowment for an outsider to produce under these circumstances. Then $Y_2(\bar{x}_1, \bar{y}_2)$ solves

$$(12) \quad \max_{y \geq 0} \{-L_2(y) + \bar{V}_2(y)\}.$$

I will say that (\bar{x}_1, \bar{y}_2) corresponds to a *competitive equilibrium* of the two-stage game if, given that the aggregate endowments will be (\bar{x}_1, \bar{y}_2) , each representative insider optimizes by producing \bar{x}_1 and each representative outsider optimizes by producing \bar{y}_2 . In the above notation

$$(13) \quad \bar{x}_1 = X_1(\bar{x}_1, \bar{y}_2)$$

and

$$\bar{y}_2 = Y_2(\bar{x}_1, \bar{y}_2).$$

This notion of competitive equilibrium requires that (\bar{x}_1, \bar{y}_2) determines a strictly monotone function $\psi(\cdot)$ in equation (8). Otherwise, there does not exist any REE in the second stage (see Theorem 1). I will say that (\bar{x}_1, \bar{y}_2) corresponds to a *candidate competitive equilibrium* if each of the prerequisites for competitive equilibrium is satisfied, except possibly the monotonicity requirement.

The following theorem is proved in the Appendix.

THEOREM 2: *The two-stage model in which insider trading is permitted has a unique candidate competitive equilibrium.*

Let aggregate endowments in the candidate equilibrium be given by (\bar{x}_1, \bar{y}_2) . If the function $\psi(\cdot)$ implied by (\bar{x}_1, \bar{y}_2) in equation (8) is monotone decreasing in β , then the two-stage model has a unique competitive equilibrium. If the function $\psi(\cdot)$ implied by (\bar{x}_1, \bar{y}_2) is not monotone, then there does not exist any competitive equilibrium.

V. Modification of the Model When Insider Trading Is Banned

In this section, I modify the two-period model that has been thus far examined in order to discuss the effects of the abstain-or-disclose requirement that have already been seen in Sections I and II. As said before, the representative insider is accorded two options in the second period: either to abstain from trading (i.e., to consume precisely his endowment brought forward from the first period) or to disclose the state before trading. I now make three additional modeling assumptions about the disclosure technology. First, any disclosure (if made) is constrained to be complete and truthful. Second, while disclosure is costless, insiders lexicographically prefer "not disclosing" to "disclosing" (all other things being equal).²⁵ Third, outsiders directly learn the true state if and only if an interval (of positive length) of insiders choose to disclose.²⁶

These additional assumptions on the disclosure technology exclude the unattractive possibility of equilibria in which insiders voluntarily disclose their information in the "insider trading permitted" regime. Without the lexicographic preference toward nondisclosure, one insider might disclose merely because one or more other insiders were also disclosing. (If he unilaterally deviated by not disclosing, there would still be other agents disclosing, and so his payoff

would be unchanged, rendering the deviation unprofitable.) With the additional assumptions, insiders do not disclose their private information unless they are required to do so in order to trade. (Whether or not one single insider chooses to disclose does not change whether an interval of positive length has disclosed and, hence, does not change the trading outcome.) At the same time, there is no inference for outsiders to draw from the fact that insiders have not disclosed their information, except that insiders were not required to disclose (cf. Grossman, 1981).

In the present model, insiders possess an independent reason for wishing to trade in the asset in question: they wish to acquire the commodity y from outsiders. (Observe that, quite generally in a model of this type, insiders will prefer to disclose rather than abstain. Under essentially any set of prices, the insider can attain strictly higher utility by trading than by abstaining from trade.) The abstain-or-disclose regulation therefore induces the informed agents to reveal their information, rendering the trading round an exchange economy with full information.

Thus, the bottom line of the subsequent analysis will be to compare a full-information equilibrium (when insider trading is banned) to that of a partially revealing REE (when insider trading is permitted) in the second round. However, the conclusion is likely to differ from that of standard comparisons of full-information versus partial-information economies, because the second-round equilibrium feeds into the determination of the first-period outcome. Given exogenous endowments (\bar{x}_1, \bar{y}_2) , insiders would typically do better when permitted to trade on their private information than in the full-information economy; they would earn trading profits at the expense of outsiders, who typically do worse when insider trading is permitted. However, since endowments are actually endogenous, the welfare comparison may be more complicated than (and different from) this straightforward result.

The analysis of the model when insider trading is banned parallels the analysis of Sections III and IV (where insider trading

²⁵This is merely the limit, as $\varepsilon \downarrow 0$, of considering a positive cost ε of disclosure.

²⁶To be precise, we assume that outsiders directly learn the true state if and only if a strictly positive measure of insiders chooses to disclose. An assumption along these lines is necessary to make the disclosure stage consistent with the notion that agents are "information takers" (i.e., that any agent should take the market information as given because his individual actions have no effect on the information available in the marketplace). This notion is implicit in competitive rational expectations equilibrium and is used throughout the paper. It would be fairly bizarre to require that a set of measure zero of insiders have absolutely no effect on the information available in the market in the trading and investment rounds, but then to allow that a set of insiders of measure zero could choose to fully inform the market in the disclosure stage.

was permitted), but the problem now becomes easier. As the abstain-or-disclose regulation induces insiders to reveal their information before they trade, outsiders now form full-information demands, which are given by

$$(6') \quad x_2^*(\phi, \bar{x}_2, \bar{y}_2; \beta, \gamma) = [w_2 / \phi] \beta$$

$$y_2^*(\phi, \bar{x}_2, \bar{y}_2; \beta, \gamma) = [w_2 / (1 - \phi)] [1 - \beta]$$

whereas insiders' demands, x_1^* and y_1^* , are still described by (3). Solving $x_1^* + x_2^* \equiv \bar{x}_1 + \bar{x}_2$ now yields

$$(8') \quad \psi^*(\beta) \equiv \frac{1 - p^*(\beta, H)}{p^*(\beta, H)}$$

$$= \frac{\bar{x}_1 \{1 - \alpha_\gamma(\beta)\} + \bar{x}_2 \{1 - \beta\}}{\bar{y}_1 \{\alpha_\gamma(\beta)\} + \bar{y}_2 \{\beta\}}$$

immediately providing a closed form for $p^*(\beta, \gamma)$. The *ex post* utility functions, V_1^* and V_2^* , are now calculated from (A2) and (A3) using p^* , x_1^* , y_1^* , x_2^* , and y_2^* . The *ex ante* expected utility functions, \bar{V}_1^* and \bar{V}_2^* , are calculated analogously as in (A4) and (A5). Optimal investment functions, $X_1^*(\cdot, \cdot)$ and $Y_2^*(\cdot, \cdot)$, can be defined analogously as in equations (11) and (12). Any $(\bar{x}_1^*, \bar{y}_2^*)$ is a competitive equilibrium if

$$(13') \quad \bar{x}_1^* = X_1^*(\bar{x}_1^*, \bar{y}_2^*)$$

and

$$\bar{y}_2^* = Y_2^*(\bar{x}_1^*, \bar{y}_2^*).$$

It is straightforward to modify the argument in the Appendix and prove that the ratio of endowments, $\bar{z}^* \equiv \bar{y}_2^* / \bar{x}_1^*$, in a competitive equilibrium corresponds to the unique fixed point of a particular mapping. It is no longer necessary to worry about whether the resulting price function is monotone, as agents are no longer required to draw any inferences from price. Existence and uniqueness of equilibrium are assured when insider trading is banned. This establishes the following theorem.

THEOREM 3: *The two-stage model in which insider trading is banned has a unique competitive equilibrium.*

VI. Welfare Implications of Insider Trading Regulation in the Model

Theorems 2 and 3 are very easy to apply to examples. While one cannot calculate closed-form solutions, the model is sufficiently simple and well-behaved that numerical approximations to the unique equilibria can be rapidly calculated on a microcomputer. Computations are best done using the reduction from endowment pairs (\bar{x}_1, \bar{y}_2) to ratios $\bar{z} \equiv \bar{y}_2 / \bar{x}_1$ developed in the Appendix.

Let $U_1(\cdot, \cdot; \cdot, \cdot), U_2(\cdot, \cdot; \cdot, \cdot), \alpha_H(\cdot), \alpha_T(\cdot)$, etc. be specified as in previous sections. Let $\omega_1 = \omega_2 \equiv \omega$ and $\rho_1 = \rho_2 \equiv \rho$, so that insiders and outsiders are assigned essentially identical disutilities of labor (in an attempt to treat them symmetrically and to avoid predetermining the conclusion).²⁷

I report now the quantitative results of an illustrative simulation and the qualitative results when the parameter values from this simulation are varied. Let $\mu_H = 2/5$ and let $\mu_T = 5/2$. Set the probability h equal to $1/2$. Meanwhile, set $\omega = 1/10$ and $\rho = 5/4$. By Theorem 2, the model in which insider trading is permitted has a unique candidate equilibrium; the ratio of endowments is calculated to equal $\bar{z} = 0.882410$. It is easily verified that the resulting price function in equations (8) and (9) is monotone, so there is indeed a unique competitive equilibrium. By Theorem 3, the model where insider trading is banned has a unique competitive equilibrium; a (simpler) computation finds that the ratio of endowments is $\bar{z}^* = 0.947958$. As described in the Appendix, these ratios immediately determine the endowment pairs under each of the two regu-

²⁷By giving insiders the same investment incentives as outsiders (as opposed to making insiders' investments inelastic) and by formulating the model so that there are "as many" insiders as outsiders, one makes it quite plausible that banning insider trading would reduce aggregate investment (since, while outsiders would invest more, insiders would seemingly invest less).

TABLE 1—INVESTMENTS AND EXPECTED UTILITIES
YIELDED BY THE SIMULATION
($\omega = 1/10$, $\rho = 5/4$, $h = 1/2$,
 $\mu_H = 2/5$, and $\mu_T = 5/2$)

INSIDER TRADING PERMITTED:	
Investment of Insiders	289.339
Investment of Outsiders	255.315
Expected Utility of Insiders	29.833
Expected Utility of Outsiders	25.514
INSIDER TRADING BANNED:	
Investment of Insiders	295.195
Investment of Outsiders	279.833
Expected Utility of Insiders	30.590
Expected Utility of Outsiders	23.613

latory regimes considered. The simulation yields the investments and expected utilities given in Table 1 [where utility now includes both $\bar{U}_i(\cdot)$ and $-L_i(\cdot)$, calculated in double precision]. The price functions that arise in the unregulated and regulated regimes are plotted in Figures 3 and 4, respectively.

The qualitative conclusions are as follows. First and most importantly, the banning of insider trading may indeed effect a Pareto improvement. In this example, it raises the representative insider's utility by 2.5 percent and raises the representative outsider's utility by 12.1 percent. Second, the principal mechanism by which this occurs is that the greater return on investment in the regulated regime induces greater investment by the outsiders. This, in turn, improves returns to insiders and actually induces greater investment by the insiders as well. Third, if the trading round had been examined in isolation, we would have instead found that the insider trading ban was merely redistributive; if agents had entered the trading stage with the (unregulated) investments of 289.339 and 255.315, but the abstain-or-disclose rule was now imposed, insiders would earn expected utilities of only 27.085, while outsiders would earn expected utilities of 31.779. Fourth, it should be noted that, in this particular model, insider trading regulation actually improves market efficiency as well; the outsiders attain full information strictly sooner.

Suppose that, starting from the parameter values of the above example, ω were to unilaterally vary anywhere in the domain

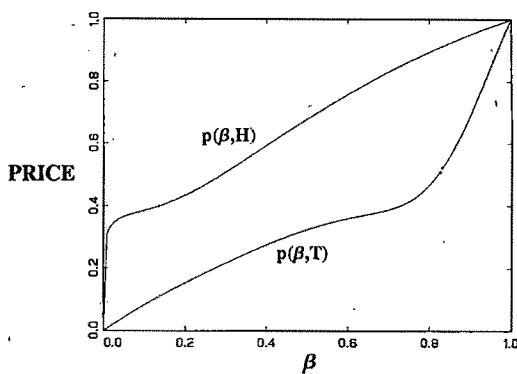


FIGURE 3. EQUILIBRIUM PRICE FUNCTION WHEN INSIDER TRADING IS PERMITTED (PLOTTED FOR $\omega = 1/10$, $\rho = 5/4$, $h = 1/2$, $\mu_H = 2/5$, AND $\mu_T = 5/2$)

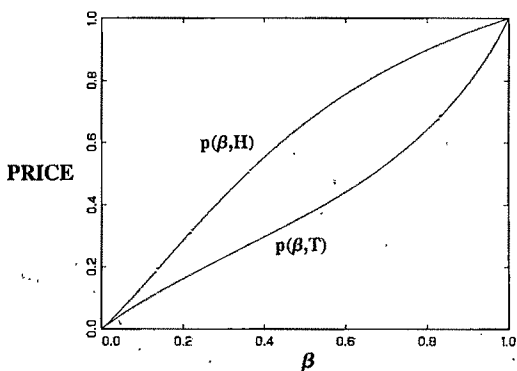
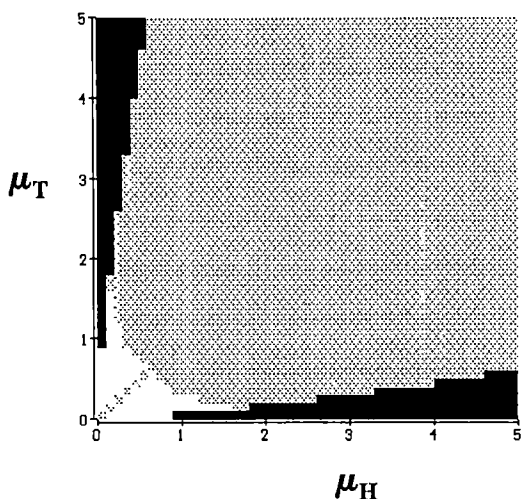


FIGURE 4. EQUILIBRIUM PRICE FUNCTION WHEN INSIDER TRADING IS BANNED (PLOTTED FOR $\omega = 1/10$, $\rho = 5/4$, $h = 1/2$, $\mu_H = 2/5$, AND $\mu_T = 5/2$)

$0 < \omega < \infty$. It is interesting that, for any such variation, the welfare implication that regulating insider trading effects a Pareto improvement is qualitatively preserved. This conclusion is also robust to unilateral perturbations of ρ anywhere in $1 < \rho < \infty$. (If $\rho \leq 1$, the strict convexity of $L_i(\cdot)$ is lost.)

A wider range of welfare conclusions is obtained by instead simultaneously perturbing the exponents μ_H and μ_T . In particular, the welfare implication that insider trading regulation effects a Pareto improvement holds only for some combinations of parameter values. For other combinations, insider trading regulation will help outsiders but






-  A Pareto improvement is effected when insider trading is banned.
-  Outsiders are helped but insiders are harmed when insider trading is banned.
-  No competitive equilibrium exists for these parameter values.

FIGURE 5. WELFARE CONSEQUENCES OF INSIDER TRADING FOR A GRID OF EXPONENTS
($\omega = 1/10$, $\rho = 5/4$, AND $h = 1/2$)

harm insiders. (Other welfare implications are logical possibilities, but I have not found any examples that lead to these possibilities.) In Figure 5, the welfare implications are calculated on a grid (of width 0.1) of values from the square $\{(\mu_H, \mu_T): 0 < \mu_H \leq 5 \text{ and } 0 < \mu_T \leq 5\}$. On most of the square, insider trading regulation indeed helps both outsiders and insiders. In a small region near the origin, outsiders are helped but insiders are harmed; however, when $\mu_H = \mu_T$, insiders and outsiders have identical preferences, the price function is the same regardless of whether insider trading is permitted or banned, and so regulation neither helps nor harms anyone. In two remaining small regions (the symmetrically arranged slivers), the last sentence of Theorem 2 comes into play: the unique candidate price function $\psi(\cdot)$ is nonmonotone, and therefore, no actual competitive equilibrium exists.

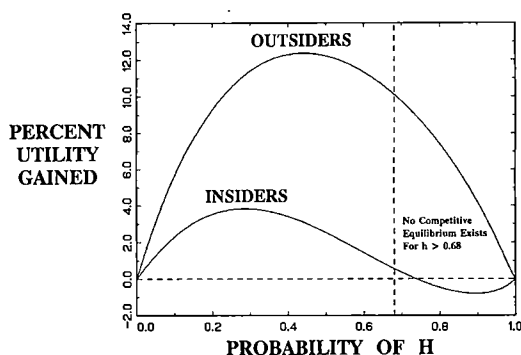


FIGURE 6. WELFARE CONSEQUENCES OF INSIDER TRADING FOR A RANGE OF PROBABILITIES
($\omega = 1/10$, $\rho = 5/4$, $\mu_H = 2/5$, AND $\mu_T = 5/2$)

Additional insight is obtained by unilaterally varying h in the interval $0 < h < 1$. One would expect that the economic consequences of changing insider trading regulations would be weakest when the informational asymmetry between insiders and outsiders is least (i.e., when h is near 0 or 1) and would be strongest when the informational asymmetry between insiders and outsiders is greatest (i.e., when h assumes intermediate values). This prediction is borne out in Figure 6, where I have perturbed the parameter h and calculated the percentage gain in *ex ante* expected utility for each class of agent. Insiders achieve the greatest percentage gain when $h \approx 0.285$, while outsiders achieve the greatest percentage gain when $h \approx 0.44$. When h is near 0 or 1, welfare effects are fairly negligible. A Pareto improvement is effected when $h \leq 0.74$; for $h > 0.74$, insiders would do better in the equilibrium in which insider trading was permitted, if one in fact existed. However, for $h > 0.68$, monotonicity of the candidate price function fails, and therefore, no competitive equilibrium exists.

If I had wished to write this article as a polemical piece, I might have emphasized another example, with $\mu_H = 1/3$, $\mu_T = 3$, $h = 1/4$, $\omega = 1/2$, and $\rho = 1.01$. With such parameter values, the effect of banning insider trading is to increase insiders' and outsiders' levels of investment each by a factor of five. Utilities of both types of agents are also increased by a factor of five.

VII. Conclusion

This article has attempted to contribute to the economic analysis of insider trading by formalizing "confidence" as an efficiency argument. If outsiders expect that insiders will take advantage of them at later stages, then outsiders may choose to invest less at the beginning. Meanwhile, effective regulation of insider trading at later stages may improve the anticipated return on investment of outsiders and, hence, promote investment by outsiders at the beginning. If insiders are helped by the availability of outside investment, insiders too may benefit from the precommitment created by insider trading regulation. It is noteworthy that the efficiency considerations posed by "confidence" point in exactly the same direction as the traditional fairness considerations, and for almost the same reason.

The confidence argument seems most likely to be decisive in scenarios where the early revelation of information affords little scope for any allocative improvement and where, in particular, the pertinent private information will become public regardless of whether trading occurs. For example, if investment bankers were permitted to trade (for their own personal benefit) shortly in advance of the announcement of tender offers, there would probably be little effect on the timing or success of tender offers. The dominant effect, *ex post*, would then be redistributive from outsiders to insiders; we should principally ask whether the anticipation of this insider trading, *ex ante*, has adverse consequences. Thus, despite the fact that my model literally made use of a commodity market rather than a stock exchange, it may still be a reasonable abstraction for assessing the desirability of insider trading in advance of tender offers.²⁸

On the other hand, my model would need to be enlarged in order to depict adequately a situation in which the revelation of private information through trading could play a

positive role in determining whether potential investment projects receive funding. Similarly, I would need to introduce additional features in order to examine a situation in which employees of an organization could be given incentives to behave as entrepreneurs.

To conclude, the simple structure of the model, which made the confidence argument relatively transparent, had the cost of allowing little other than the confidence effect to occur. A valuable next step in assessing the desirability of insider trading would be the construction of richer models that have room for both the negative incentives posed by confidence and the positive incentives described by previous authors. Such analysis may eventually give us a better understanding of the specific types of contexts in which insider trading should be permitted or banned.

APPENDIX

Derivation of the Conditional Probability $\pi(\beta)$ Used by Outsiders. Suppose the representative outsider knew that $\phi_0 \leq p(\tilde{\beta}, \tilde{\gamma}) \leq \phi_1$, where $\phi_0 = p(\beta_0, H)$ and $\phi_1 = p(\beta_1, H)$. He could infer that either $\tilde{\gamma} = H$, in which case $\beta_0 \leq \beta \leq \beta_1$, or else $\tilde{\gamma} = T$, in which case $\alpha(\beta_0) \leq \beta \leq \alpha(\beta_1)$. Since $\tilde{\beta}$ and $\tilde{\gamma}$ are independent, observe that the unconditional probability of the first event (Ω_H in Fig. 2) is $h[\beta_1 - \beta_0]$ and the unconditional probability of the second event (Ω_T in Fig. 2) is $[1 - h][\alpha(\beta_1) - \alpha(\beta_0)]$. Hence, the conditional probability of the first event is given by

$$(A1) \quad \Pr[\tilde{\gamma} = H | \phi_0 \leq p(\tilde{\beta}, \tilde{\gamma}) \leq \phi_1] \\ = \frac{h[\beta_1 - \beta_0]}{h[\beta_1 - \beta_0] + [1 - h][\alpha(\beta_1) - \alpha(\beta_0)]}.$$

The probability that $(\tilde{\beta}, \tilde{\gamma}) = (\beta, H)$, conditional on $p(\tilde{\beta}, \tilde{\gamma}) = p(\beta, H)$, is calculated by setting $\beta_0 = \beta$ and taking the limit as $\beta_1 \rightarrow \beta_0$ of the expression in (A1). Using l'Hôpital's rule, this yields equation (4).

Derivation of the Agents' Ex Ante Expected Utility Functions. Let $V_1(x; \beta, \gamma)$ ($V_2(y; \beta, \gamma)$)

²⁸Hence, this article should be viewed as lending support to Rule 14e-3 of the SEC, which in a very expansive way regulates trading in advance of tender offers (see also Section I).

denote the *ex post* utility attained by a representative insider (outsider) who has carried forward x (y) units of endowment, when the actual state in period two is (β, γ) . Insiders and outsiders apply demands from equations (3) and (6), respectively, giving attained utilities of

$$(A2) \quad V_1(x; \beta, \gamma) \\ = x \{ U_1 [x_1(p(\beta, \gamma), 1, 0; \beta, \gamma), \\ y_1(p(\beta, \gamma), 1, 0; \beta, \gamma); \beta, \gamma)] \}$$

and

$$(A3) \quad V_2(y; \beta, \gamma) \\ = y \{ U_2 [x_2(p(\beta, \gamma), 0, 1; \beta, \gamma), \\ y_2(p(\beta, \gamma), 0, 1; \beta, \gamma); \beta, \gamma)] \}.$$

In the above equations, the terms x and y factor out in a linear fashion because of the homogeneity of degree one of the functions $U_i(\cdot)$, $x_i(\cdot)$, and $y_i(\cdot)$. The *ex ante* expected utilities, $\tilde{V}_1(x)$ and $\tilde{V}_2(y)$, are calculated by merely integrating the *ex post* utilities over all possible states, using the correct unconditional probabilities. Using (A2) and (A3), one obtains

$$(A4) \quad \tilde{V}_1(x) = x \tilde{V}_1(1) \\ = x \left\{ h \int_0^1 V_1(1; \beta, H) d\beta \right. \\ \left. + (1-h) \int_0^1 V_1(1; \beta, T) d\beta \right\}$$

and

$$(A5) \quad \tilde{V}_2(y) = y \tilde{V}_2(1) \\ = y \left\{ h \int_0^1 V_2(1; \beta, H) d\beta \right. \\ \left. + (1-h) \int_0^1 V_2(1; \beta, T) d\beta \right\}.$$

PROOF OF THEOREM 2:

The rational expectations equilibrium price function, as indicated in equations (8) and (9), is homogeneous of degree zero in aggregate endowment (\bar{x}_1, \bar{y}_2) ; that is, if aggregate endowments equaled (\bar{x}_1', \bar{y}_2') such that $\bar{y}_2'/\bar{x}_1' = \bar{y}_2/\bar{x}_1$, precisely the same price function would result. This follows from the fact that the utility functions [of (1) and (2)] imply demands [given by (3) and (6)] which themselves are homogeneous of degree one in endowments. Similarly, observe that $X_1(\cdot, \cdot)$ and $Y_2(\cdot, \cdot)$, defined in (11) and (12), are homogeneous functions of degree zero.

Since the important features of the model depend only on the ratio of aggregate endowments, define $\bar{z} \equiv \bar{y}_2/\bar{x}_1$ to be the ratio of endowments. Also, define the mapping $T(\bar{z})$ to yield the optimal ratio of endowments for insiders and outsiders to produce individually in the first period if they expect the ratio of aggregate endowments in the second period to equal \bar{z} . Utilizing the homogeneity of degree zero, one sees that $T(\cdot)$ is given by

$$(A6) \quad T(\bar{z}) \equiv Y_2(1, \bar{z})/X_1(1, \bar{z}).$$

I will now establish the existence and uniqueness of a fixed point of $T(\cdot)$.

Consider the trading stage as $\bar{z} \rightarrow \infty$. Note, using (8) and (9), that the price function $p(\cdot, \cdot)$ of good x converges to one, pointwise. Therefore, a representative insider with endowment $\bar{x}_1 = 1$ can afford to purchase an arbitrarily large quantity of good y while still consuming $x_1 = 1/2$, implying $\tilde{V}_1(1) \rightarrow \infty$. Meanwhile, a representative outsider with endowment $\bar{y}_2 = 1$ can barely afford to purchase any quantity of good x , implying $\tilde{V}_2(1) \rightarrow 0$. Extracting the first-order conditions from equations (11) and (12) and substituting from equations (A4) and (A5) yields

$$(A7) \quad X_1(1, \bar{z}) = \{ \tilde{V}_1(1) / \rho_1 \omega_1 \}^{1/(\rho_1 - 1)}$$

$$Y_2(1, \bar{z}) = \{ \tilde{V}_2(1) / \rho_2 \omega_2 \}^{1/(\rho_2 - 1)}$$

and leads to the conclusion that $X_1(1, \bar{z}) \rightarrow \infty$ and $Y_2(1, \bar{z}) \rightarrow 0$. The definition of $T(\cdot)$ in (A6) thus implies $\lim_{\bar{z} \rightarrow \infty} T(\bar{z}) = 0$.

Consider the trading stage as $\bar{z} \rightarrow 0$. Analogous reasoning yields that $\bar{V}_1(1) \rightarrow 0$ and, hence, $X_1(1, \bar{z}) \rightarrow 0$; similarly, $\bar{V}_2(1) \rightarrow \infty$ and, hence, $Y_2(1, \bar{z}) \rightarrow \infty$. Thus, $\lim_{\bar{z} \rightarrow 0} T(\bar{z}) = \infty$. This demonstrates that there exist z^1 and z^2 , where $0 < z^1 < z^2$, such that $T(z^1) > z^1$ and $T(z^2) < z^2$. Since the mapping $T(\cdot)$ is continuous, the intermediate-value theorem guarantees the existence of a fixed point \bar{z} between z^1 and z^2 .

The uniqueness of a fixed point is established by demonstrating that $T(\cdot)$ is monotone decreasing. Consider any two ratios of endowment, \bar{z} and \bar{z}' , where $\bar{z}' > \bar{z} > 0$. Let $p(\cdot, \cdot)$ and $p'(\cdot, \cdot)$ be the price functions implied by \bar{z} and \bar{z}' , respectively, using equations (8) and (9). Note, for any given state (β, γ) , where $0 < \beta < 1$, that $p'(\beta, \gamma) > p(\beta, \gamma)$. Now observe that when the price is $p'(\beta, \gamma)$, the representative outsider's demands, $x_2(p'(\beta, \gamma), 0, 1; \beta, \gamma)$ and $y_2(p'(\beta, \gamma), 0, 1; \beta, \gamma)$, are also within the outsider's budget constraint (with some slack in wealth remaining) at the lower price $p(\beta, \gamma)$. Hence, $V_2(1; \beta, \gamma)$, defined in equation (A3), is strictly greater at price $p(\beta, \gamma)$ than at price $p'(\beta, \gamma)$. Since this inequality holds whenever $0 < \beta < 1$, the *ex ante* expected utility $\bar{V}_2(1; \beta, \gamma)$, defined in equation (A5), is also strictly greater at price $p(\beta, \gamma)$ than at price $p'(\beta, \gamma)$. Using (A7), this establishes that $Y_2(1, \bar{z}) > Y_2(1, \bar{z}')$. Analogous reasoning for the representative insider establishes that $X_1(1, \bar{z}') > X_1(1, \bar{z})$. Using (A6), one concludes that $T(\bar{z}') < T(\bar{z})$. Now suppose there existed two fixed points \bar{z} and \bar{z}' , where $\bar{z}' > \bar{z}$. This would imply $T(\bar{z}') - T(\bar{z}) = \bar{z}' - \bar{z}$, contradicting that $T(\cdot)$ is monotone decreasing. Therefore, $T(\cdot)$ has a unique fixed point \bar{z} .

Finally, observe that there is a one-to-one correspondence between candidate competitive equilibria of the two-stage game and fixed points of $T(\cdot)$. The aggregate endowments (\bar{x}_1, \bar{y}_2) associated with a candidate equilibrium imply a fixed point by $\bar{z} = \bar{y}_2 / \bar{x}_1$; a fixed point \bar{z} yields aggregate endowments of a candidate equilibrium by $\bar{x}_1 = X_1(1, \bar{z})$ and $\bar{y}_2 = Y_2(1, \bar{z})$. This leads

to the conclusion that there exists a unique candidate competitive equilibrium, which is also an actual competitive equilibrium if and only if $\psi(\cdot)$ is monotone.

REFERENCES

- Allen, Beth, "Generic Existence of Completely Revealing Equilibria for Economies with Uncertainty when Prices Convey Information," *Econometrica*, September 1981, 49, 1173-99.
- Ausubel, Lawrence M., "Partially-Revealing Rational Expectations Equilibrium in a Competitive Economy," *Journal of Economic Theory*, February 1990, 50, 93-126.
- Bagwell, Laurie S., "Share Repurchase and Takeover Deterrence," Northwestern University, Department of Finance, Working Paper No. 53, July 1988.
- , "Dutch Auction Repurchases: An Analysis of Shareholder Heterogeneity," Northwestern University, mimeo, August 1989.
- and Judd, Kenneth L., "Transaction Costs and Corporate Control," Northwestern University, mimeo, November 1988.
- Balcer, Yves and Judd, Kenneth L., "Effects of Capital Gains Taxation on Life-Cycle Investment and Portfolio Management," *Journal of Finance*, July 1987, 42, 743-61.
- Benabou, Roland and Laroque, Guy, "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility," Massachusetts Institute of Technology, mimeo, February 1989.
- Bhattacharya, Uptal and Spiegel, Matthew, "Insiders, Outsiders and Market Breakdowns," Columbia University, mimeo, October 1989.
- Brudney, Victor, "Insiders, Outsiders, and Informational Advantages Under the Federal Securities Laws," *Harvard Law Review*, December 1979, 93, 322-76.
- Carlton, Dennis W. and Fischel, Daniel R., "The Regulation of Insider Trading," *Stanford Law Review*, May 1983, 35, 857-95.
- Dennert, Jurgen, "Insider Trading and the Allocation of Risks," Universitat Basel, mimeo, October 1989.
- Diamond, Douglas W. and Verrecchia, Robert E.,

- "Information Aggregation in a Noisy Rational Expectations Economy," *Journal of Financial Economics*, September 1981, 9, 221-35.
- Dye, Ronald A., "Inside Trading and Incentives," *Journal of Business*, July 1984, 57, 295-313.
- Easterbrook, Frank H., "Insider Trading, Secret Agents, Evidentiary Privileges, and the Production of Information," *Supreme Court Review*, 1981, 11, 309-65.
- Fishman, Michael J. and Hagerty, Kathleen M., "Insider Trading and the Efficiency of Stock Prices," Northwestern University, Department of Finance, mimeo, April 1989.
- Gale, Douglas and Hellwig, Martin F., "Informed Speculation in Large Markets," University of Pittsburgh, mimeo, August 1987.
- Grossman, Sanford J., "The Informational Role of Warranties and Private Disclosure about Product Quality," *Journal of Law and Economics*, December 1981, 24, 461-83.
- _____, and Stiglitz, Joseph E., "On the Impossibility of Informationally Efficient Markets," *American Economic Review*, June 1980, 70, 393-408.
- Haddock, David D. and Macey, Jonathan R., "Regulation on Demand: A Private Interest Model, with an Application to Insider Trading Regulation," *Journal of Law and Economics*, October 1987, 30, 311-52.
- Jaffe, Jeffrey F., "Special Information and Insider Trading," *Journal of Business*, July 1974, 47, 410-28.
- Jordan, James S., "On the Efficient Markets Hypothesis," *Econometrica*, September 1983, 51, 1325-44.
- Kihlstrom, Richard E. and Postlewaite, Andrew, "Equilibrium in a Securities Market with a Dominant Trader Possessing Inside Information," University of Pennsylvania, mimeo, June 1983.
- Kyle, Albert S., "Continuous Auctions and Insider Trading," *Econometrica*, November 1985, 53, 1315-35.
- Laffont, Jean-Jacques and Maskin, Eric S., "The Efficient Market Hypothesis and Insider Trading on the Stock Market," *Journal of Political Economy*, February 1990, 98, 70-93.
- Langevoort, Donald C., *Insider Trading Regulation*, New York: Clark Boardman, 1990.
- Lorie, James H. and Niederhoffer, Victor, "Predictive and Statistical Properties of Insider Trading," *Journal of Law and Economics*, April 1968, 11, 35-53.
- Manne, Henry G., (1966a) *Insider Trading and the Stock Market*, New York: Free Press, 1966.
- _____, (1966b) "In Defense of Insider Trading," *Harvard Business Review*, November-December 1966, 44, 113-22.
- Manove, Michael, "The Harm from Insider Trading and Informed Speculation," *Quarterly Journal of Economics*, November 1989, 104, 823-45.
- Posner, Richard A., *Economic Analysis of Law*, Boston: Little, Brown & Co., 1986.
- Radner, Roy, "Rational Expectations Equilibrium: Generic Existence and the Information Revealed by Prices," *Econometrica*, May 1979, 47, 655-78.
- Scott, Kenneth E., "Insider Trading: Rule 10b-5, Disclosure, and Corporate Privacy," *Journal of Legal Studies*, December 1980, 9, 801-18.
- Seligman, Joel, "The Reformulation of Federal Securities Law Concerning Nonpublic Information," *Georgetown Law Journal*, April 1985, 73, 1083-1140.
- Seyhun, H. Nejat, "Insiders' Profits, Costs of Trading, and Market Efficiency," *Journal of Financial Economics*, June 1986, 16, 189-212.
- Shleifer, Andrei, "Do Demand Curves for Stocks Slope Down?," *Journal of Finance*, July 1986, 41, 579-90.

Optimal Bypass and Cream Skimming

By JEAN-JACQUES LAFFONT AND JEAN TIROLE*

This paper develops a normative model of regulatory policy toward bypass and cream skimming. It analyzes the effects of bypass on second-degree price discrimination, on the rent of the regulated firm, and on the welfare of low-demand customers. It shows that pricing under marginal cost may be optimal for the regulated firm, excessive cream skimming occurs if access to the bypass technology is not regulated, and the prohibition of bypass may increase or decrease the regulated firm's rent. (JEL 026, 613)

The regulation of "natural monopolies" is often associated with policies toward competition, including restrictions on entry. This paper is concerned with a common form of competition, which threatens the regulated firm on its most lucrative markets. Examples abound: in the telecommunications industry, the development of microwave radio and communication satellites in the 1960's introduced the possibility that big telecommunication customers might bypass the major common carrier (AT&T) and deal directly with a satellite company, for instance. More recently, some large firms have bypassed the local telephone networks and have acquired direct links to long-distance carriers. Similar issues arise in the energy sector. Big industrial consumers of electricity may generate their own power; and the 1978 Natural Gas Policy Act in the United States has created the possibility for industrial plants to bypass the local distribution utilities by building direct connections to the pipelines, gas producers, or intermediaries.

What distinguishes these examples from other situations in which a regulated firm

faces competition is that the competitive pressure focuses on the high-demand customers (the "cream") and not on low-demand ones (the "skimmed milk"). That is, entry interferes with second-degree price discrimination by the regulated monopolist.¹

As one would expect, cream-skimming has been the object of much regulatory attention. Regulated monopolies have repeatedly called for entry restrictions. For instance, AT&T has assailed MCI as a cream skimmer, lapping up the profits on favorable routes and eschewing high-cost low-return service, and has accused Comsat of syphoning the most profitable part of the business. More recently, local distribution companies have made similar charges against bypass. In both cases, the regulated monopoly has argued that, because of economies of scale, bypass would raise the rates of small-volume commercial and residential users or would reduce the quality of their service. Historically, the case for restriction of entry into the market of a natural monopoly has often been supported on such cream-skimming

*GREMAQ, Université de Toulouse, France, and Massachusetts Institute of Technology, respectively. The authors are grateful to the Ford Foundation, the Pew Charitable Trust, the Guggenheim Foundation, the Center for Energy Policy Research at MIT, the National Science Foundation, and the French Ministère de l'Éducation Nationale for financial support. Helpful comments on a previous draft were supplied by David Sappington and Richard Schmalensee.

¹Cream skimming is also often discussed in the context of a multiproduct firm when some of the regulated firm's most valuable products are skimmed off by competitors. The case of third-degree price discrimination is simpler to analyze than that of second-degree discrimination, because different customers are offered different terms and therefore there are no incentive constraints of consumers to satisfy; see our previous paper (Laffont and Tirole, 1990b) for an analysis of some issues concerning the interaction of a multiproduct firm with its competitors.

grounds. Conversely, some have held the view that bypass is the outcome of a healthy competition and can only result in efficiency gains when it occurs. For instance, the Federal Energy Regulatory Commission has argued against the denial of certificates to competitors on the basis that local distribution companies are in a position to compete aggressively.²

This paper develops a normative model of regulatory policy toward bypass and cream skimming. It posits a double asymmetry of information. First, the regulated firm is ignorant of the demand characteristics of individual customers and must thus practice second-degree price discrimination.³ There are two types of customers: "high-demand" and "low-demand." The high-demand customers have the opportunity of using an alternative, competitively supplied technology. The fixed cost associated with this alternative technology (e.g., the cost of building an interconnection or a generator) makes bypass unattractive to low-demand customers. Second, the regulated firm knows more about its own technology than the regulator. The firm's rent is affected by the policy toward bypass. Although the cost function is assumed to be linear in total output, it exhibits increasing returns to scale as a bigger output makes reductions in marginal cost more desirable. The regulator chooses the pricing, cost reimbursement, and possibly bypass policies to maximize social welfare.

The paper's technical contribution is twofold. First, the theory of nonlinear pricing has focused on "downward binding" incentive constraints; that is, a monopolist must design a pricing scheme that prevents

high-demand customers from consuming the low-demand customers' bundle (see Eric Maskin and John Riley, 1984; Michael Mussa and Sherwin Rosen, 1978). In the presence of bypass, the monopolist may have to offer advantageous terms to high-demand customers in order to retain them. This may lead low-demand customers to consume the high-demand customers' bundle, even though they would not use the bypass technology. The paper studies the effect of "upward binding" incentive constraints.⁴ Second, the bypass technology introduces discontinuities in the control of the regulated firm. We show how to deal with such discontinuities and extend results obtained in regulatory models with more conventional net-demand functions, in particular the linearity of optimal cost-reimbursement rules (see Laffont and Tirole, 1986).

The economic contribution is a welfare analysis of cream skimming. We ask: a) whether asymmetric information between the regulator and the regulated firm increases the amount of bypass; b) how, in a situation of asymmetric information, the regulator can ask the firm to substantiate its claim that bypass should either be prohibited or prevented through price cuts; c) whether a marginal price under marginal cost is an appropriate response to the threat of bypass; d) whether low-demand customers are hurt by the possibility of bypass; e) whether there is socially too much or too little bypass; and f) whether the regulated firm is necessarily hurt by the possibility of bypass. Section I describes the model, Section II derives the optimal regulatory scheme, and Section III summarizes our main findings.⁵

²See Alfred Kahn (1971 chapters 1,4,6) for a discussion of the economic arguments in favor of and against cream skimming.

³Imperfect information about consumers may justify cross-subsidization for redistributive purposes (see Laffont and Tirole, 1990a). A normative analysis of entry in such circumstances could be carried out along the lines of this paper. Preventing entry may facilitate redistribution in the same way it may facilitate second-degree price discrimination here.

⁴For similar considerations, see Paul Champsaur and Jean-Charles Rochet (1989) for a duopoly model of competition in quality and prices, as well as Tracy Lewis and David Sappington (1989) for countervailing incentives in regulation. Incentive constraints may also be "binding upwards" (but for another reason) in dynamic incentive problems without commitment (see Laffont and Tirole, 1987).

⁵Bernard Caillaud (1985) studies the effect of an unregulated competitive fringe on the regulation of a

I. The Model

The regulated firm serves two types of consumers ($i = 1, 2$), in numbers α_1 and α_2 . Let q_1 and q_2 be the consumptions of type-1 and type-2 consumers, respectively. Total consumption is $Q = \alpha_1 q_1 + \alpha_2 q_2$. Let $S_i(q_i)$ be the utility derived by a type- i consumer from consuming the regulated firm's good. To facilitate the analysis,⁶ we make the assumptions that $S_1(q) = S(q)$ and $S_2(q) = \theta S(q)$ with $\theta > 1$.

The technology of the regulated firm is defined by its cost function:

$$(1) \quad C = (\beta - e)Q$$

where β is an intrinsic cost parameter known only to the firm, and e is an effort level which has disutility $\psi(e)$ (with $\psi' > 0$, $\psi'' > 0$, $\psi''' \geq 0$)⁷ for the firm's manager. We could add a (known) fixed cost in (1) without any change for our analysis. It is common knowledge that $\beta \in [\underline{\beta}, \bar{\beta}]$.

The regulator observes the firm's outputs q_1 and q_2 , cost C , and revenue $R(q_1, q_2)$. We make the accounting convention that the regulator receives $R(q_1, q_2)$, reimburses cost C , and pays a net transfer t to the firm.

natural monopoly. As in this paper, the possibility of substitution for consumers introduces discontinuities in the regulated firm's control problem. Caillaud focuses on the role of correlation between the technologies of the regulated firm and the competitive fringe, rather than on the issue of second-degree price discrimination (in Caillaud's model, arbitrage constrains firms to practice linear pricing). Michael Einhorn (1987) provides an analysis of bypass in a model without asymmetric information about the firm's technology. He obtains the result that marginal price may be below marginal cost for some consumers in the absence of incentive constraints for consumers. In our analysis, price may be below marginal cost because the regulated firm cannot identify high-valuation consumers and use third-degree price discrimination. Einhorn (1987) finds that customers make efficient choices when deciding to use the bypass technology. We will show, on the contrary, that bypass is used too often, compared to a situation in which the regulator could monitor the access to bypass.

⁶This assumption enables us to have a convex program in the no-bypass regimes so that necessary and sufficient conditions for characterizing the optimal solution are available.

⁷ $\psi''' \geq 0$ makes stochastic incentive schemes nonoptimal.

Regulation must satisfy the individual rationality (IR) constraint of the firm:⁸

$$(2) \quad U = t - \psi(e) \geq 0 \quad \forall \beta \in [\underline{\beta}, \bar{\beta}].$$

In addition, there exists an alternative (bypass) technology to which a consumer has access if he pays a fixed cost f . Then, the constant marginal cost of this alternative technology is d .^{9,10}

We make an assumption that ensures that type-1 consumers (the low-valuation consumers) never find it advantageous to use the bypass technology. Let

$$S_1^* \equiv \max_q [S(q) - f - dq]$$

denote the utility level of a type-1 consumer using the bypass technology. We postulate that the bypass alternative is never an optimal choice for low-demand consumers: $S_1^* \leq 0$. We assume that α_1/α_2 is not too small, so that it is always optimal for the regulated firm to serve the low-demand customers. However, in some circumstances, the type-2 (high-valuation) consumers may want to quit the regulated firm and use the bypass technology.

To fit the examples given in the introduction, we assume that individual consumption of the regulated product can be monitored by the regulated firm so that nonlinear pricing is feasible. Let T_i denote the payment by type- i consumers for quantity q_i and let $\{(T_1, q_1), (T_2, q_2)\}$ be a nonlinear schedule. Then, $R(q_1, q_2) = \alpha_1 T_1 + \alpha_2 T_2$.

The regulator is utilitarian and wishes to maximize the sum of consumers' welfares

⁸We assume implicitly that it is never optimal to shut down the regulated firm.

⁹Note that our analysis would be unchanged if the regulated firm could also offer the alternative technology and Bertrand competition with zero profits on the high-demand consumers took place.

¹⁰Our approach thus differs from that of the contestability literature (William Baumol et al., 1982) in several respects. First, we allow transfers between the regulator and the firm. Second, the regulated firm and its competitors (here, the bypass producers) do not face the same cost functions. Third, the regulated firm's technology is unknown to the regulator.

and the firm's utility level, taking into account that the social cost of public funds is $1 + \lambda > 1$ (because of distortive taxation). When all consumers consume the regulated firm's good, social welfare is

$$\begin{aligned}
 (3) \quad W &= \alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\
 &\quad - (\alpha_1 T_1 + \alpha_2 T_2) - (1 + \lambda) \\
 &\quad \times (C + t - \alpha_1 T_1 - \alpha_2 T_2) + U \\
 &= \alpha_1 S(q_1) + \alpha_2 \theta S(q_2) - (1 + \lambda) \\
 &\quad \times [(\beta - e)(\alpha_1 q_1 + \alpha_2 q_2) + \psi(e)] \\
 &\quad - \lambda U + \lambda(\alpha_1 T_1 + \alpha_2 T_2).
 \end{aligned}$$

When type-2 consumers use the bypass technology, their utility level is $S_2^* = \max_q (\theta S(q) - f - dq)$, and social welfare is

$$\begin{aligned}
 (4) \quad W^b &= \alpha_1 S(q_1) + \alpha_2 S_2^* - (1 + \lambda) \\
 &\quad \times [(\beta - e)\alpha_1 q_1 + \psi(e)] \\
 &\quad - \lambda U + \lambda \alpha_1 T_1.
 \end{aligned}$$

The transfers T_1 and T_2 will be seen to be linear combinations of $S(q_1)$ and $S(q_2)$; so if we denote $S(q_1) \equiv s_1$ and $S(q_2) \equiv s_2$, the concavity of the objective function W relies on the concavity in (s_1, s_2, e) of $\Gamma(s_1, s_2, e) = - (1 + \lambda) \{ (\beta - e) [\alpha_1 \zeta(s_1) + \alpha_2 \zeta(s_2)] + \psi(e) \}$ where ζ is the inverse function of S . Throughout the paper we assume that $\Gamma(\cdot)$ is strictly concave in the relevant domain of (s_1, s_2, e) .¹¹ This will ensure a unique solution for (q_1, q_2, e) in each regime considered in the paper.

The regulator does not observe e and has incomplete information about β . He has a prior on $[\beta, \bar{\beta}]$ represented by the cumulative distribution function $F(\beta)$ which satisfies the monotone hazard-rate property $(d/d\beta)(F/f) > 0$.¹² The first step of our

analysis is to characterize the regulator's optimal pricing rule and optimal incentive schemes under incomplete information.

II. Optimal Pricing Rule and Optimal Incentive Schemes

Our first observation is that incentive compatibility implies that the marginal cost $c(\beta)$ is a nondecreasing function of the intrinsic cost parameter β . Moreover, the particular cost function we postulated for the regulated firm makes the rent of asymmetric information $U(\beta)$ that must be given up to a firm of type β a function only of the marginal cost schedule $c(\cdot)$ being implemented. These facts, which follow from classical arguments in incentive theory, are proved in Appendix 1. The intuition as to why the firm's rent depends only on the marginal cost schedule $c(\cdot)$ can be grasped from the cost function (1). The regulator observes C and Q and, therefore, knows the realized marginal cost. Thus, the scope for the firm to transfer good technological conditions (low β) into a slack rent (low e) are determined by the marginal cost schedule. The intuition as to why c is nondecreasing in β is that it is less costly for a low- β firm to produce at a low cost. Hence, if type β prefers to produce at cost c rather than cost $\bar{c} < c$, type $\beta' > \beta$ cannot prefer to produce at cost \bar{c} rather than cost c .

These two facts justify the two-step procedure that we use to characterize the optimal solution. For a given value of β and a given average (or marginal) cost c , social welfare is

$$\begin{aligned}
 (5) \quad W(c, q_1, q_2, T_1, T_2, \beta) \\
 = V(c, q_1, q_2, T_1, T_2) + Z(\beta, c)
 \end{aligned}$$

where

$$\begin{aligned}
 Z &\equiv - (1 + \lambda) \psi(\beta - c) - \lambda U(\beta) \\
 V &\equiv \alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\
 &\quad - (1 + \lambda) c(\alpha_1 q_1 + \alpha_2 q_2) \\
 &\quad + \lambda(\alpha_1 T_1 + \alpha_2 T_2)
 \end{aligned}$$

¹¹ If we assume that $\psi(e) \rightarrow \infty$, as $e \rightarrow \bar{e}$ and $\beta - \bar{e} > 0$ for all β , concavity is obtained if S'' is large enough.

¹² This assumption simplifies the analysis by preventing "bunching phenomena" (bunching occurs when the regulator induces different types β to choose the same allocation $\{t, c, q_1, q_2\}$).

when the bypass technology is not used by type-2 consumers, and

$$V \equiv \alpha_1 S(q_1) + \alpha_2 S_2^* - (1 + \lambda) c \alpha_1 q_1 + \lambda \alpha_1 T_1$$

when it is.

The constraints imposed on the regulator's maximization program are of two kinds: the firm's incentive constraints

$$(6) \quad \dot{U}(\beta) = -\psi'(\beta - c(\beta)) \quad \text{and}$$

$$\dot{c}(\beta) \geq 0$$

$$(7) \quad U(\bar{\beta}) \geq 0$$

almost everywhere (see Appendix 1) and the consumers' incentive constraints coming from the fact that the firm cannot distinguish *ex ante* between the two types of consumers. Type-1 consumers' individual rationality constraint (IR₁) is

$$(8) \quad S(q_1) - T_1 \geq 0.$$

Type-2 consumers must obtain a utility level as large as in the bypass alternative to remain with the regulated firm; therefore, their individual rationality constraint (IR₂) is

$$(9) \quad \theta S(q_2) - T_2 \geq S_2^*.$$

When the type-2 consumers use the bypass technology, IR₂ becomes irrelevant.

Self-selection by consumers also imposes the following incentive constraints (IC₁ and IC₂, respectively):

$$(10) \quad S(q_1) - T_1 \geq S(q_2) - T_2$$

$$(11) \quad \theta S(q_2) - T_2 \geq \theta S(q_1) - T_1.$$

In view of the decomposition obtained in (5), the optimization of expected social welfare under constraints (6)–(11) can be decomposed into 1) a maximization of V subject to (8)–(11) with respect to q_1 , q_2 , T_1 , and T_2 for each value of c (which determines optimal pricing) and 2) a maximization of the expected value of social welfare with respect to $c(\cdot)$ for those values of q_1 ,

TABLE 1—THE SIX POSSIBLE REGIMES

Regime	Binding Constraints ^a				Bypass?
1	(IC ₂)	(IR ₁)			No
2	(IC ₂)	(IR ₁)	(IR ₂)		No
3		(IR ₁)	(IR ₂)		No
4		(IR ₁)	(IR ₂)	(IC ₁)	No
5			(IR ₂)	(IC ₁)	No
6		(IR ₁)			Yes

^aAs given in equations (8)–(11).

q_2 , T_1 , and T_2 under constraints (6) and (7).

Let us now consider the first maximization. Straightforward arguments show that constraints (8)–(11) define six possible regimes characterized by the binding constraints in (8)–(11) and the occurrence of the bypass regime (see Appendix 2 for a proof). These regimes are given in Table 1.

In each of the six regimes, optimal pricing is determined by

$$\max[V(c, q_1, q_2, T_1, T_2)]$$

subject to the binding constraints of regime i .

Let $\tilde{V}^i(c)$ be the optimal value of this program in regime i and $q_j^i(c)$ the consumption of type j in regime i , $i = 1, \dots, 6$. The results of these maximizations are gathered in the next proposition. Let $p_1 \equiv S'(q_1)$ and $p_2 \equiv \theta S'(q_2)$ denote the marginal prices for the type-1 and type-2 consumers, respectively.

PROPOSITION 1: *In regimes 1 and 2, $p_1 > c$ and $p_2 = c$ with $dp_1/dc > 0$ in regime 1 and $dp_1/dc = 0$ in regime 2. In regime 3, $p_1 = p_2 = c$. In regimes 4 and 5, $p_1 = c$ and $p_2 < c$ with $dp_2/dc = 0$ in regime 4 and $dp_2/dc > 0$ in regime 5. In regime 6, $p_1 = c$.*

See Appendix 3 for a proof of this proposition and for the formulas defining optimal prices. In Proposition 2 (below), we show that regimes are ordered from 1 through 6 as c [or equivalently β from (6)] increases.

For a very efficient regulated firm (very low β), the net surplus obtained by high-valuation consumers is strictly higher than what they could obtain with the bypass, even when

the firm uses optimal second-degree price discrimination. The classical "no distortion at the top" result amounts to the equality of marginal price and marginal cost for the high-valuation consumers (regime 1). For low-valuation consumers the marginal price exceeds marginal cost to prevent high-valuation consumers from buying the low-valuation buyers' bundle (because the shadow cost of public funds is positive, the regulated firm behaves qualitatively like a monopolist and thus tries to limit the rent enjoyed by high-valuation consumers). As the efficiency of the regulated firm decreases, the bypass constraint becomes binding. The allocation is then distorted in several steps. The payment T_2 that can be obtained from the high-valuation consumers must be limited, which relaxes IC_2 and thus allows the regulator to bring the low-valuation consumers' marginal price closer to marginal cost (regime 2). As β still increases, IC_2 becomes nonbinding, and the regulator equates marginal cost and marginal price for both types (regime 3); but as β still increases, the limit put on the payment made by type-2 consumers makes type-1 consumers' incentive constraint (IC_1) binding. They wish to take the contract offered to type 2. To prevent that, consumption of type-2 consumers is increased beyond the first best level by lowering the marginal price below marginal cost; this relaxes IC_1 , because type-2 consumers have a higher marginal utility for the good (regime 4).¹³ Finally, the payment made by the type-1 consumers is lowered to satisfy their incentive constraint, which leaves them with a surplus. When this regime is obtained, we have the interesting result that, because of (unmonitored) bypass, optimal regulation may require leaving a rent to low-valuation

consumers to be able to offer to high-valuation consumers a deal good enough that they do not use the bypass (regime 5). Last, in the bypass regime (regime 6), the firm serves a single category of consumers and thus imposes no distortion in consumption. Note that, while the optimal regime depends on β , all the conclusions about pricing in a given regime obtained in Proposition 1 hold for any β (i.e., are due only to the asymmetry of information with respect to consumers' tastes).

We next observe that the bypass regime can only occur for an interval $[\beta^*, \bar{\beta}]$ of values of β ,¹⁴ and more generally the regimes (when they exist) are ordered from 1 to 6 when β increases.

PROPOSITION 2: (i) *There exists $\beta^* \in [\beta, \bar{\beta}]$, such that the bypass regime occurs if and only if $\beta > \beta^*$.* (ii) *There exist $\{\beta_i\}_{i=0, \dots, 6}$ with $\beta_i \leq \beta_{i+1}$, $\beta_0 = \beta$, $\beta_5 = \beta^*$, $\beta_6 = \bar{\beta}$ such that regime i prevails if and only if $\beta \in [\beta_{i-1}, \beta_i]$.*

See Appendix 4 for a proof.

The intuition for (i) is simply that, if the regulated firm becomes more inefficient, letting the high-demand customers use the bypass becomes more attractive. The idea behind the proof of (ii) is that, given the concavity of our problem, variables that satisfy the first-order conditions for maximizing expected social welfare and yield continuous control variables form a solution. Moving from regime 1 to regime 5 yields a continuous solution on $[\beta, \beta^*]$. Figures 1 and 2 summarize our findings up to now.

We next look for a solution to the maximization over $[\beta, \beta^*]$ with the ordering of regimes 1, 2, 3, 4, and 5 along the β -axis. Let us call $\tilde{V}(c(\beta))$ the function obtained by piecing together the functions $\tilde{V}^1(c), \dots, \tilde{V}^5(c)$ on the intervals $[\beta, \beta_1]$, $[\beta_1, \beta_2], \dots, [\beta_4, \beta^*]$, with $\beta \leq \beta_1 \leq \beta_2 \leq \beta_3$

¹³For a given marginal cost c , the consumptions of low-demand customers in regime 2 (q_1^2) and high-demand customers in regime 4 (q_2^4) are equal. This is due to the facts that $\theta S(q_1^2) - T_1 = S^*$ (from IC_2 and IR_2) and $T_1 = S(q_1^2)$ (from IR_1) on the one hand, and $\theta S(q_2^4) - T_2 = S^*$ (from IR_2) and $T_2 = S(q_2^4)$ (from IC_1 and IR_1) on the other hand yield the same solution: $q_1^2 = q_2^4 = S^{-1}(S^*/(\theta - 1))$. (This is not an artifact of the particular surplus functions we chose.)

¹⁴This observation is of interest only when $\bar{\beta}$ is sufficiently large. For small $\bar{\beta}$, $\beta^* = \bar{\beta}$ and bypass never occurs (although it may be binding).

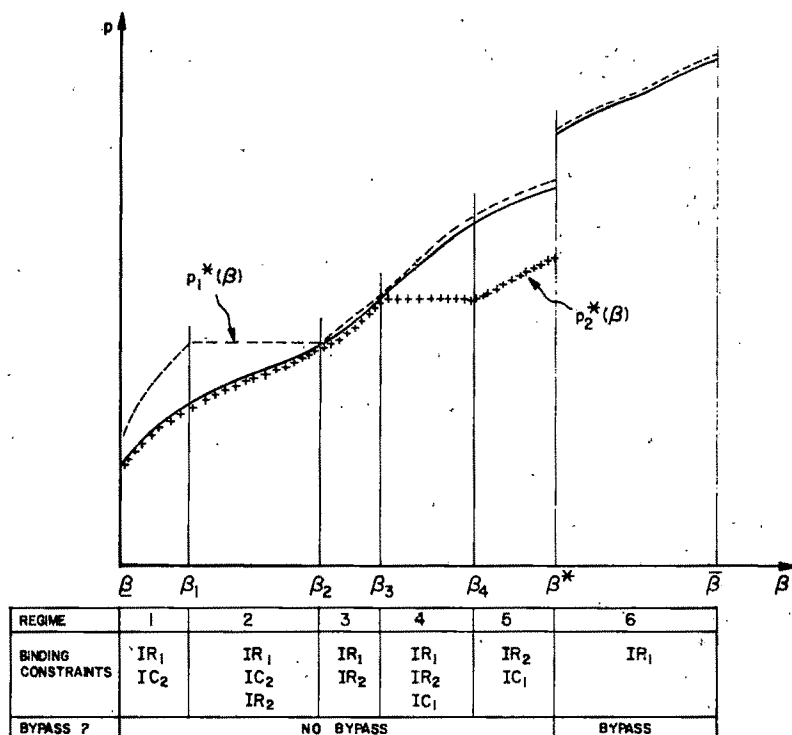
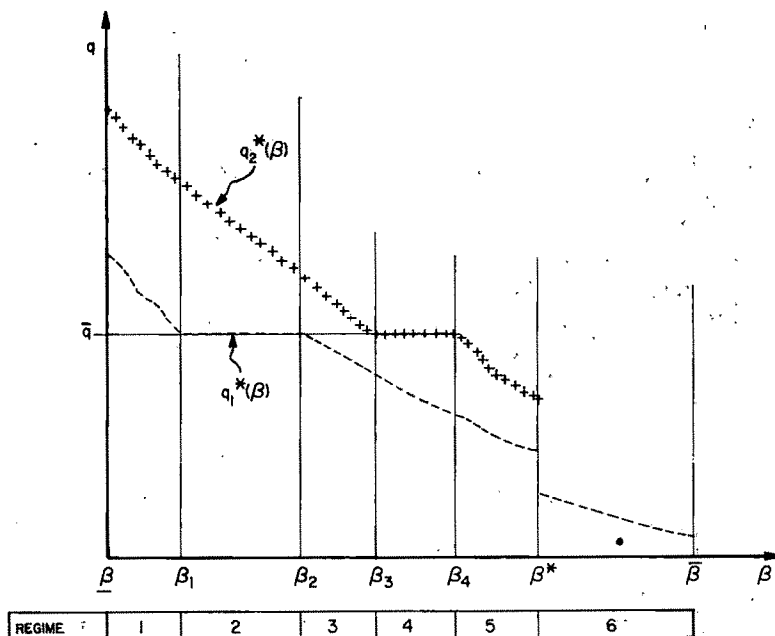
FIGURE 1. PRICE PROFILES [SOLID LINE = MARGINAL COST = $\beta - e^*(\beta)$]

FIGURE 2. QUANTITY PROFILES

$\leq \beta_4 \leq \beta^* \leq \bar{\beta}$ (all regimes need not exist). That is, $\tilde{V}(c) = \max_{i \in \{1, \dots, 5\}} [\tilde{V}^i(c)]$.

In view of Proposition 2, the program for the overall maximization of expected welfare can be written:

$$(12) \quad \max_{\underline{\beta}} \int_{\underline{\beta}}^{\beta^*} [\tilde{V}(c(\beta)) - (1 + \lambda) \times \psi(\beta - c(\beta)) - \lambda U(\beta)] dF(\beta) \\ + \int_{\beta^*}^{\bar{\beta}} [\tilde{V}^6(c^6(\beta)) - (1 + \lambda) \times \psi(\beta - c^6(\beta)) - \lambda U(\beta)] dF(\beta)$$

subject to

$$\dot{U}(\beta) = -\psi'(\beta - c(\beta))$$

$$U(\bar{\beta}) \geq 0$$

$$\dot{c}(\beta) \geq 0$$

(this last constraint will be ignored in a first step and checked *ex post*).

Fixing β^* and $U(\beta^*) = \bar{U} = \int_{\beta^*}^{\bar{\beta}} \psi'(\bar{\beta} - c^6(\bar{\beta})) d\bar{\beta}$, we first look for a solution to the maximization over $[\underline{\beta}, \beta^*]$ using the fact that the ordering of regimes is 1, 2, 3, 4, 5 along the β -axis. We have

$$(13) \quad \max \sum_{i=0}^4 \int_{\beta_i}^{\beta_{i+1}} \tilde{V}(c(\beta)) dF(\beta) \\ - \int_{\underline{\beta}}^{\beta^*} [(1 + \lambda)\psi(\beta - c(\beta)) \\ + \lambda U(\beta)] dF(\beta)$$

subject to

$$\dot{U}(\beta) = -\psi'(\beta - c(\beta))$$

$$U(\beta^*) = \bar{U}.$$

In each regime, using $d\tilde{V}^i/dc = -(1 + \lambda)(\alpha_1 q_1^i + \alpha_2 q_2^i)$, and letting $q_j^i(c)$ denote the optimal consumption of type- j consumers in regime i when marginal cost is c (see

Proposition 1), the Pontryagin principle yields

$$(14) \quad \psi'(\beta - c^{i^*}(\beta)) \\ = \alpha_1 q_1^{i^*}(c^{i^*}(\beta)) + \alpha_2 q_2^{i^*}(c^{i^*}(\beta)) \\ - \left[\frac{\lambda}{1 + \lambda} \right] \left[\frac{F(\beta)}{f(\beta)} \right] \psi''(\beta - c^{i^*}(\beta))$$

as in Laffont and Tirole (1986). Equation (14) has a unique solution in each regime from our concavity assumptions.

The characterizations of the (possibly degenerate) intervals defining regimes are obtained by maximizing (13) with respect to $\beta_1, \beta_2, \beta_3$, and β_4 :

$$(15) \quad \tilde{V}^i(c^{i^*}(\beta_i)) = \tilde{V}^{i+1}(c^{i+1}(\beta_i)) \\ \text{for } i \in \{1, 2, 3, 4\}.$$

Furthermore, the derivative of the value of the program (13) with respect to $\bar{U} = U(\beta^*)$ is equal to $\lambda F(\beta^*)$ (a unit increase in \bar{U} translates into a unit increase in the rent of all types that are more efficient than β^* , which has social cost λ).

Still fixing β^* , we next look for a solution to

$$(16) \quad \max_{\beta^*} \int_{\beta^*}^{\bar{\beta}} [\tilde{V}^6(c(\beta)) - (1 + \lambda) \times \psi(\beta - c(\beta)) - \lambda U(\beta)] dF(\beta) \\ - \lambda F(\beta^*) U(\beta^*)$$

with the constraints that

$$\dot{U}(\beta) = -\psi'(\beta - c(\beta))$$

$$U(\bar{\beta}) \geq 0.$$

Since $d\tilde{V}_1^6/dc = -(1 + \lambda)q_1^6(c)$, the Pontryagin principle gives

$$(17) \quad \psi'(\beta - c^6(\beta)) = \alpha_1 q_1^6(c^6(\beta)) \\ - \frac{\lambda}{1 + \lambda} \left(\frac{F(\beta)}{f(\beta)} \right) \\ \times \psi''(\beta - c^6(\beta))$$

with

$$U(\beta^*) = \int_{\beta^*}^{\bar{\beta}} \psi'(\tilde{\beta} - c^{6*}(\tilde{\beta})) d\tilde{\beta}.$$

Using the expression for q_1^6 given in Proposition 1, our concavity assumptions of Section I imply that (17) has a unique solution $\{q_1^{6*}(\beta), c^{6*}(\beta)\}$.

The rents are obtained by backward induction from regime 6 to regime 1:

$$(18) \quad U(\beta) = \int_{\beta}^{\bar{\beta}} \psi'(\tilde{\beta} - c^{6*}(\tilde{\beta})) d\tilde{\beta} \quad \text{for } \beta \in [\beta^*, \bar{\beta}]$$

$$U(\beta) = \int_{\beta}^{\beta^*} \psi'(\tilde{\beta} - c^{5*}(\tilde{\beta})) d\tilde{\beta} + U(\beta^*) \quad \text{for } \beta \in [\beta_4, \beta^*]$$

...

$$U(\beta) = \int_{\beta}^{\beta_1} \psi'(\tilde{\beta} - c^{1*}(\tilde{\beta})) d\tilde{\beta} + U(\beta_1) \quad \text{for } \beta \in [\underline{\beta}, \beta_1].$$

It remains to optimize with respect to β^* . For λ small enough, the problem is strictly concave in β^* (Appendix 5). Assuming that regime $i \in \{1, \dots, 5\}$ is to the left of β^* , we obtain the first-order equation (using the continuity of the rent)

$$(19) \quad \tilde{V}^i(c^i(\beta^*)) - (1 + \lambda)\psi(\beta^* - c^i(\beta^*)) \\ = \tilde{V}^6(c^6(\beta^*)) - (1 + \lambda)\psi(\beta^* - c^6(\beta^*)).$$

(Figures 1 and 2 depict the case in which all five regimes exist to the left of β^* .) Since the objective function is concave in β^* for all i , there exists a unique solution. Which regime i prevails before bypass occurs depends on the values of the parameters (see Appendix 6).

Remark on Two-Part Tariffs: A similar analysis can be performed when the firm is constrained to charging a two-part tariff: $T(q)$

$= a + bq$.¹⁵ The number of relevant regimes is lower, as there are no longer incentive constraints for the two types of customers (each type of customer chooses his preferred point on the straight-line tariff). So IR_1 alone, IR_2 alone, or both individual rationality constraints¹⁶ may be binding. Results similar to those for nonlinear pricing can then be obtained. For instance, if only the bypass constraint (IR_2) is binding (which can be shown to be optimal for some values of the parameters), the optimal slope of the two-part tariff is given by

$$[b - (\beta - e)] \left(\alpha_1 \frac{dq_1}{db} + \alpha_2 \frac{dq_2}{db} \right) \\ = \alpha_1 [q_2(b) - q_1(b)] \frac{\lambda}{1 + \lambda}.$$

Because $q_2(b) > q_1(b)$ and $dq_i/db < 0$ for all b and i , $b < \beta - e$; the regulated firm's marginal price is below marginal cost.

Finally, we will show that the optimal transfer schedule can be implemented through a menu of linear contracts. Since there is no atom at β^* , the β^* firm is indifferent between the bypass regime and the no-bypass regime. From Laffont and Tirole (1986), we know that in the bypass and no-bypass regions the nonlinear transfer schedule $t(c)$ is convex.¹⁷ It is therefore also globally convex across regimes (see Fig. 3).

Therefore, this schedule can be replaced by a menu of linear contracts with slope

¹⁵We maintain our assumption that the regulator chooses an incentive scheme $t(c)$ for the firm. As before, the firm's revenue is not necessarily raised entirely by the direct charges to final consumers.

¹⁶In the latter case, a and b , and therefore $q_1(b)$ and $q_2(b)$, are completely determined by the two constraints.

¹⁷The convexity of $t(c)$ in $[c(\beta), c(\beta^*)]$ results directly from Laffont and Tirole (1986). The proof must be slightly extended in $[\beta^*, \bar{\beta}]$, because of the term $\lambda F(\beta^*)U(\beta^*)$ in program (16), which represents the shadow cost of the rent at the left of the interval; but the shadow cost and, therefore, the allocation on $[\beta^*, \bar{\beta}]$ are the same as if regime 6 obtained on $[\beta, \bar{\beta}]$. Therefore, again from Laffont and Tirole (1986), $t(c)$ is convex on $[c(\beta^*), c(\bar{\beta})]$.

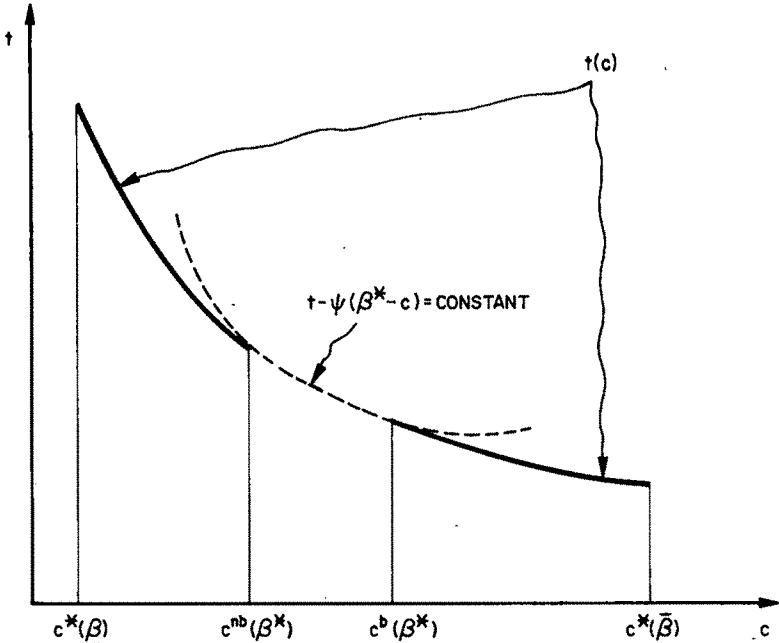


FIGURE 3. IMPLEMENTATION THROUGH LINEAR SCHEMES (c = MARGINAL COST; NB = NO BYPASS; B = BYPASS)

$\psi'(\beta(c) - c)$, where $\beta(c)$ is the type that produces at marginal cost c . The interest of this result is that an additive noise can be added in the cost function (1) without any effect on our results.

III. Bypass and Cream Skimming

In this section, we compare the optimal regulatory mechanism characterized in Section II with the one obtained when, in addition, the regulator can monitor the access to the bypass technology. We assume that the regulator's new instrument is the possibility of prohibiting bypass (that is, bypass is only partially regulated). The regulator's objective function is unchanged, but the constraints are now

(22) $S(q_1(\beta)) - T_1(\beta) \geq 0$

(23) $\theta S(q_2(\beta)) - T_2(\beta) \geq 0$

when the bypass technology is not used, and only (22) when the bypass technology is used by high-valuation consumers.

Then, as in traditional adverse-selection problems, only the high-valuation incentive constraint and the low-valuation individual rationality constraint are binding. For $\beta \leq \beta^{c^*}$, we are therefore in regime 1 of Section II. For $\beta \geq \beta^{c^*}$ we are in regime 6 of Section II. The pricing policy is illustrated in Figure 4. Optimization with respect to the value β^{c^*} , at which bypass starts being allowed, yields

(20) $\theta S(q_2(\beta)) - T_2(\beta) \geq \theta S(q_1(\beta)) - T_1(\beta)$

(21) $S(q_1(\beta)) - T_1(\beta) \geq S(q_2(\beta)) - T_2(\beta)$

(24) $\tilde{V}^1(c^1(\beta^{c^*})) - (1 + \lambda)\psi(\beta^{c^*} - c^1(\beta^{c^*})) = \tilde{V}^6(c^6(\beta^{c^*})) - (1 + \lambda)\psi(\beta^{c^*} - c^6(\beta^{c^*}))$

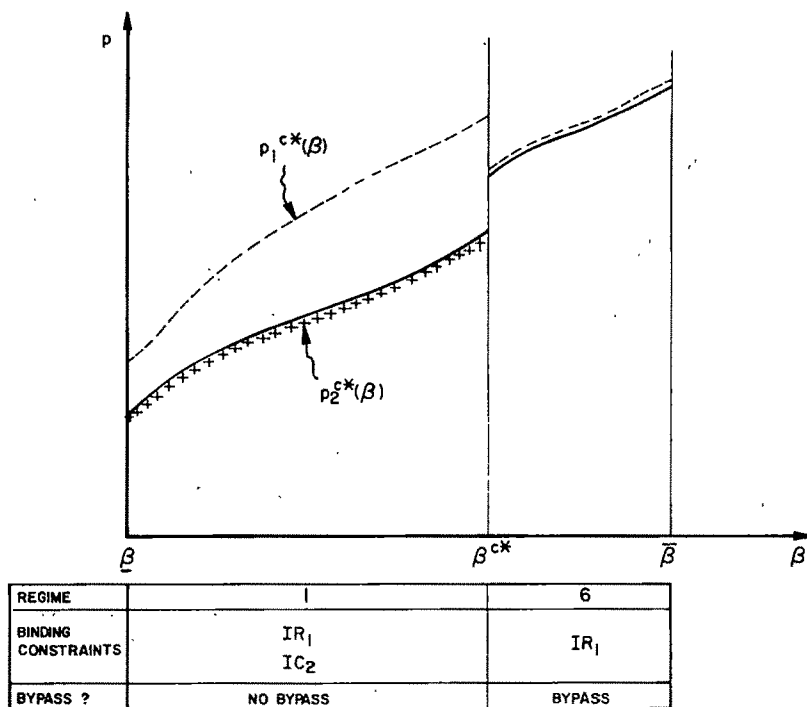


FIGURE 4. PRICE PROFILES WHEN BYPASS IS MONITORED (SOLID LINE = MARGINAL COST)

(the superscript c identifies this case of "control" of access to bypass).

PROPOSITION 3: $\beta^{c*} \geq \beta^*$; there is excessive bypass when bypass cannot be prohibited.

PROOF:

Suppose that $\beta^{c*} < \beta^*$. From the definition of β^* , we have

$$\begin{aligned}
 (25) \quad & \int_{\beta^{c*}}^{\beta^*} [\tilde{V}(\tilde{c}(\beta)) - (1 + \lambda) \\
 & \quad \times \psi(\beta - \tilde{c}(\beta)) - \lambda \tilde{U}(\beta)] f(\beta) d\beta \\
 & \quad - \lambda F(\beta^{c*}) \tilde{U}(\beta^{c*}) \\
 & \geq \int_{\beta^{c*}}^{\beta^*} [V^6(c^6(\beta)) - (1 + \lambda) \\
 & \quad \times \psi(\beta - c^6(\beta)) \\
 & \quad - \lambda U^6(\beta)] f(\beta) d\beta \\
 & \quad - \lambda F(\beta^{c*}) U^6(\beta^{c*})
 \end{aligned}$$

where

$$(26) \quad U^6(\beta) \equiv \int_{\beta}^{\tilde{\beta}} \psi'(\tilde{\beta} - c^6(\tilde{\beta})) d\tilde{\beta}$$

$$\begin{aligned}
 (27) \quad & \tilde{U}(\beta) \equiv \int_{\beta}^{\beta^*} \psi'(\tilde{\beta} - \tilde{c}(\tilde{\beta})) d\tilde{\beta} \\
 & \quad + U^6(\beta^*)
 \end{aligned}$$

and where \tilde{c} and \tilde{V} refer to c^i and \tilde{V}^i for the optimal regime $i \in \{1, \dots, 5\}$ (as in Section II). In words, when bypass cannot be prohibited, the regulator could have used regime 6 on $[\beta^{c*}, \beta^*]$ but elected not to do so. Now (25) is satisfied a fortiori if $\tilde{V}(\tilde{c}(\beta))$ is replaced by $V^1(\tilde{c}(\beta))$ because $V^1(c) \geq \tilde{V}(c)$ for all c (there are fewer constraints when bypass can be prohibited). This means that when bypass can be prohibited, the regulator would be better off prohibiting bypass on $[\beta^{c*}, \beta^*]$ even if he chose the suboptimal function $\tilde{c}(\cdot)$ instead of $c^1(\cdot)$ on that interval, a contradiction.

The intuition for Proposition 3 is that when the regulated firm supplies high-demand consumers, the threat of bypass imposes an additional constraint that reduces welfare relative to the one (V^1) that can be obtained when bypass is prohibited. Because the prohibition of bypass eliminates this constraint and raises welfare in the no-bypass region, it becomes optimal to extend the latter region.

Next we compare the firm's rents $U(\beta)$ and $U^c(\beta)$ when bypass cannot and can be prohibited.

PROPOSITION 4: *There exists $\beta_0 \geq \underline{\beta}$ such that*

$$U^c(\beta) < U(\beta) \quad \text{for } \beta < \beta_0,$$

$$U^c(\beta) > U(\beta) \quad \text{for } \beta_0 < \beta < \beta^*,$$

$$U^c(\beta) = U(\beta) \quad \text{for } \beta \geq \beta^*.$$

Furthermore, $\beta_0 \geq \beta_1$ if $\beta_0 > \underline{\beta}$.

PROOF:

The rent in both cases is given by $\int_{\underline{\beta}}^{\beta} \psi'(e(\tilde{\beta})) d\tilde{\beta}$, where

$$(28) \quad \psi'(e(\tilde{\beta})) = Q(\tilde{\beta}) - \frac{\lambda}{1+\lambda} \times \frac{F(\tilde{\beta})}{f(\tilde{\beta})} \psi''(e(\tilde{\beta})).$$

Next we note that $Q^i(c) \geq Q^1(c) > Q^6(c)$ for $i \in \{1, \dots, 5\}$ and for all c .¹⁸ This results from Proposition 1 (or Fig. 1); in regimes 2 and 3, q_2 is the same as in regime 1 (for the given marginal cost c), while q_1 is higher. In regimes 4 and 5, both q_1 and q_2 are higher than in regime 1.

¹⁸Note the parallel between bypass and the "shut-down option" in traditional adverse-selection models (e.g., David Baron and Roger Myerson, 1982). In the shutdown regions, the firm's rent does not increase with its efficiency. Here, it increases at a slower rate in the bypass region ($Q^i > Q^6$).

Because in all regimes

$$(29) \quad \psi'(\beta - c^i(\beta)) = Q^i(c^i(\beta)) - \frac{\lambda}{1+\lambda} \times \frac{F(\beta)}{f(\beta)} \psi''(\beta - c^i(\beta))$$

and because the objective function is concave, we get $c^i(\beta) \leq c^1(\beta)$ for all $i \in \{1, \dots, 5\}$. Hence $Q^i(\beta) \equiv Q^i(c^i(\beta)) \geq Q^i(c^1(\beta)) \geq Q^1(c^1(\beta)) \equiv Q^1(\beta)$.

Equation (28) thus implies that $\psi'(e(\beta))$ (which is also the slope of the incentive scheme at β) is higher on $[\beta_1, \beta^*]$ and smaller on $[\beta^*, \beta^*]$ when bypass cannot be prohibited. The slopes coincide on $[\beta, \beta_1]$ and on $[\beta^*, \beta]$ (see Fig. 5). Proposition 4 follows immediately.

The intuition for Proposition 4 is as follows. When bypass can be prohibited, the bypass region is smaller. Thus, in $[\beta^*, \beta^*]$, where bypass is now avoided, the regulated firm supplies the high-demand customers, which raises demand and makes marginal cost reduction more desirable. Thus, more incentives are given to the firm to reduce costs, which raises the rent of firms in $[\beta^*, \beta^*]$. However, in former regimes 2–5 (if such regimes exist), keeping the high-demand customers no longer requires the high outputs characterized in Proposition 1, as bypass can be prohibited. The regulator reduces the incentives for cost reduction and thus the firm's rent. That is, as β decreases under β^* , $U^c(\beta) - U(\beta)$ decreases and may become negative. The firm need not gain from the prohibition of bypass, because the threat of bypass was a "good excuse" for low prices, high outputs, and thus a high rent.

Let us next consider the effect of a change of information on the extent of bypass when access to bypass is controlled. We index the distribution $F(\beta, \nu)$ and the inverse of the hazard rate $H(\beta, \nu) \equiv F(\beta, \nu)/f(\beta, \nu)$ by a parameter ν . We assume that $H_\nu < 0$ (that is, the hazard rate increases with ν). For some families of distributions, an increase in the hazard rate corresponds to an improvement in information. For instance, for a uniform distribution on $[\underline{\beta}, \beta]$, $H = \beta - \underline{\beta}$,

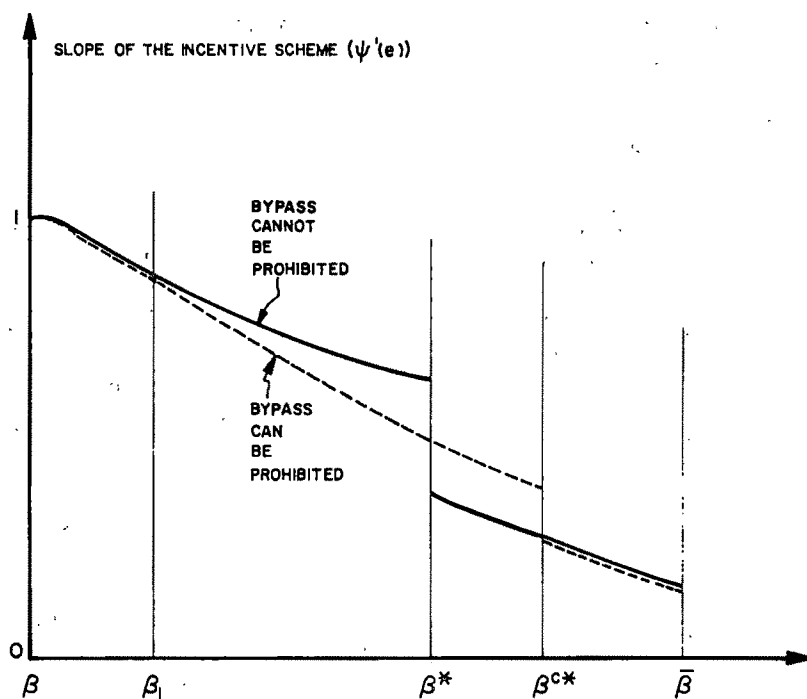


FIGURE 5. SLOPE OF THE INCENTIVE SCHEME

so that when β increases, the support of the uniform distribution shrinks (in this example, $\nu = \beta$). Differentiating (24) we get

$$(30) \quad \frac{d\beta^{c*}}{d\nu} = \frac{\lambda}{1+\lambda} \left(\frac{H(\beta, \nu)}{Q^1 - Q^6} \right) \times \left(\frac{d}{d\nu} [\psi'(e^1(\beta^{c*}))] - \frac{d}{d\nu} [\psi'(e^6(\beta^{c*}))] \right)$$

Intuitively, an increase of ν , creates more bypass if, at β^{c*} , the rate of increase of the (costly) rent ψ' is less affected in the no-bypass regime than in the bypass regime.

Differentiating the first-order conditions defining quantities and effort [see (28) as well as regimes 1 and 6 in Appendix 3], we

obtain

$$(31) \quad \frac{de^1}{d\nu} = -H_\nu \left[\frac{\lambda}{1+\lambda} \right] \times \left[\frac{\psi''(e^1)}{\psi''(e^1) + \frac{\lambda}{1+\lambda} \psi'''(e^1)H + \frac{dQ^1}{dc}} \right]$$

$$(32) \quad \frac{de^6}{d\nu} = -H_\nu \left[\frac{\lambda}{1+\lambda} \right] \times \left[\frac{\psi''(e^6)}{\psi''(e^6) + \frac{\lambda}{1+\lambda} \psi'''(e^6)H + \frac{dQ^6}{dc}} \right]$$

PROPOSITION 5: Assume that ψ'' is constant and that either λ is small or demand functions are concave. Then an increase in ν

that increases the hazard rate f/F around β^c increases β^c .¹⁹

PROOF:

We know from Appendix 4 that $Q^1 > Q^6$. Hence, $e^1 > e^6$, from (14) and the associated second-order condition. Next, λ small or concave demand functions imply that $dQ^1/dc < dQ^6/dc < 0$ (see the expressions in Appendix 3). Thus (12), (13), and (14) imply that $d\beta^c/d\nu > 0$.

The intuition for Proposition 5 is that, as f/F increases, the concern with the firm's rent in the no-bypass region (which has probability F) decreases relative to the concern for the distortion at β^c (which consists in imposing bypass to reduce the rent of better types and has probability f). The need for assumptions in Proposition 5 comes from the fact that output, and not only the firm's rent and cost, matters. Without such assumptions the result might be reversed.

For the case in which bypass cannot be prohibited, the result is similar and true even more often, since the slope $|dQ/dc|$ is higher in that case (at least for regimes 3, 4, and 5).

Finally, let us note that, at least for small λ , bypass increases with λ , both with and without control of access to bypass. Intuitively, more-costly transfers make bypass more desirable by increasing the incentive costs of the regulated firm. This holds at least for small λ ; for large λ , the social gain stemming from the firm's revenue $\lambda R(q)$ may upset this result. We now prove this result in the case where bypass is controlled, but a similar reasoning holds in the other case.

PROPOSITION 6: *Bypass increases with λ , for λ small.*

PROOF:

Let $SC(\beta) \equiv (\beta - e)Q + \psi(e)$ denote total social cost, and let $R(\beta) \equiv \alpha_1 T_1 + \alpha_2 T_2$

denote total revenue. Differentiating (24) gives

$$(33) \quad \left. \frac{d\beta^c}{d\lambda} \right|_{\lambda=0} = \frac{[R^1(\beta^c) - SC^1(\beta^c)] - [R^6(\beta^c) - SC^6(\beta^c)]}{Q^1(\beta^c) - Q^6(\beta^c)}.$$

At $\lambda = 0$,

$$R^1(\beta^c) - SC^1(\beta^c) = W^1(\beta^c) - (\theta - 1) \times \alpha_2 S(q_1^1(\beta^c))$$

$$R^6(\beta^c) - SC^6(\beta^c) = W^6(\beta^c) + \alpha_2 S_2^*.$$

As social welfare is continuous at β^c , the result follows since $Q^1(\beta^c) > Q^6(\beta^c)$.

IV. Conclusion

Our main economic findings may be summarized as follows:

- Asymmetric information between the regulator and the firm raises the actual cost of the regulated firm and increases the probability of bypass.²⁰
- Bypass should be fought (i.e., high-demand customers should be retained) when the regulated firm is efficient. An efficient firm is screened through its choice of a steep (high-slope) incentive scheme. Hence, the slope of the regulated firm's incentive scheme is positively correlated with its success in fighting bypass.
- It may be optimal to charge marginal prices below marginal cost for high-demand customers. Because these customers must be granted advantageous terms to be retained, low-demand customers must be dissuaded from buying the high-demand customers' bundle by charging a high fixed fee and a low marginal price.

¹⁹The probability of bypass is equal to $1 - F(\beta^c, \nu)$. The total effect of an increase in ν is in general ambiguous, as F decreases with ν but increases with β^c , which itself increases with ν .

²⁰The role of bypass in providing discipline for the regulated firm is similar to the role played by entry and auditing in Joel Demski et al. (1987) and in David Scharfstein (1988).

- (d) Low-demand customers are not necessarily hurt by the threat of bypass. In our model, they may enjoy a positive net consumer surplus when the regulated firm is constrained by bypass but does not let bypass operate, while they never do when bypass is controlled by the regulator. They indirectly benefit from the high-demand customers' being offered advantageous terms. With more than two customer types, it can be shown that some customers may be hurt by bypass.²¹ Our point here is that the effect of bypass on low-demand customers is ambiguous; the "skimmed milk" need not be made worse off by bypass, contrary to conventional wisdom.
- (e) There is excessive bypass if bypass cannot be controlled by the regulator. Bypass interferes with optimal second-degree price discrimination.
- (f) A mediocre regulated firm is hurt by bypass. An efficient regulated firm may benefit from the threat of bypass, because it can use it to vindicate high levels of production.

Caution should be exercised in particular when applying the last two conclusions. We compared two regulatory institutions, in which the regulator has or does not have the authority to prohibit the competitive technology. The analysis is restrictive for two (related) reasons. First, in principle there exists a vast array of government interventions with regard to competition, which include direct regulation and subsidies or taxation. Second, we did not make explicit the reasons why the government has limited authority on the competitive sector. Presumably, this limitation in scope of authority stems from the costs of regulation or from the fear that the extension of the scope of regulatory authority from the dominant firm to the whole industry would result in producer protection. Despite these

caveats, we believe that a normative analysis such as the one performed here is a first step toward understanding the policy trade-offs with regard to bypass and cream skimming.

Caution should also be exercised in the study of particular industries. While we analyzed optimal regulation, transfers from the regulators to the firms are sometimes legally prohibited (e.g., in the telecommunications and electricity industries in the United States). A regulated firm's cost is then entirely paid by direct charges to consumers. Some of our conclusions may be affected by the impossibility of transfers; see Laffont and Tirole (1990c) for an attempt at explaining the prohibition of transfers. For instance, low-demand consumers are more likely to be hurt by bypass in the presence of returns to scale when consumers must pay the firm's full cost.

Last, several additional issues could be addressed within our normative framework. For instance, if the regulated firm were to choose *ex ante* among technologies, would it prefer a high-investment low-marginal-cost technology to fight bypass or a low-investment high-marginal-cost one to focus on low-demand customers, and how would this affect high- and low-demand customers? Also, we have ignored some potential benefits of competition. If the bypass technology is similar to the regulated firm's, the regulator can use bypass as a yardstick to monitor the firm further. Moreover, it might be the case that bypass enhances product variety by making available goods that cannot be produced by the regulated firm. These and other questions are left open for future research.

APPENDIX 1

From the revelation principle, a regulatory scheme can be represented by a revelation mechanism which specifies for each announcement of the cost characteristic $\hat{\beta}$ levels of production $q_1(\hat{\beta})$ for a type-1 consumer and $q_2(\hat{\beta})$ for a type-2 consumer, a total cost target $C(\hat{\beta})$, and a net transfer received by the firm from the regulator $t(\hat{\beta})$.

²¹ For instance, with three types, if type-3 (high-demand) customers use the bypass technology, the costs of producing for the remaining two types increase because of returns to scale, which may reduce the net consumer surplus of type-2 customers.

Incentive compatibility at β and β' requires

$$\begin{aligned} (A1) \quad & \hat{U}(\beta, \beta) \\ &= t(\beta) - \psi\left(\beta - \frac{C(\beta)}{\alpha_1 q_1(\beta) + \alpha_2 q_2(\beta)}\right) \\ &\geq \hat{U}(\beta, \beta') \\ &= t(\beta') - \psi\left(\beta - \frac{C(\beta')}{\alpha_1 q_1(\beta') + \alpha_2 q_2(\beta')}\right) \end{aligned}$$

$$\begin{aligned} (A2) \quad & \hat{U}(\beta', \beta') \\ &= t(\beta') - \psi\left(\beta' - \frac{C(\beta')}{\alpha_1 q_1(\beta') + \alpha_2 q_2(\beta')}\right) \\ &\geq \hat{U}(\beta', \beta) \\ &= t(\beta) - \psi\left(\beta' - \frac{C(\beta)}{\alpha_1 q_1(\beta) + \alpha_2 q_2(\beta)}\right) \end{aligned}$$

Adding (A1) and (A2) and denoting $c(\beta) \equiv C(\beta)/[\alpha_1 q_1(\beta) + \alpha_2 q_2(\beta)]$ (the average cost), we get

$$\begin{aligned} (A3) \quad & \psi(\beta' - c(\beta)) - \psi(\beta - c(\beta)) \\ &\geq \psi(\beta' - c(\beta')) - \psi(\beta - c(\beta')). \end{aligned}$$

Consider the function: $\Phi(x) \equiv \psi(\beta' - x) - \psi(\beta - x)$. For $\beta < \beta'$, $\Phi'(x) < 0$ since $\psi'' > 0$. Therefore, (A3) implies $c(\beta) \leq c(\beta')$. The revelation mechanism can alternatively be represented by the functions $q_1(\cdot)$, $q_2(\cdot)$, $c(\cdot)$, and $t(\cdot)$.

Let $U(\beta)$ be the rent captured by a firm of type β . Incentive compatibility implies that $U(\beta)$ is continuous and nonincreasing, since a type- β firm with $\beta < \beta'$ can always mimic a type- β' firm at smaller cost. $U(\beta)$ is therefore almost everywhere differentiable. Similarly, $c(\beta)$ is almost everywhere differentiable. $\dot{U}(\beta)$ exists almost everywhere and

$$\dot{U}(\beta) = -\psi'(\beta - c(\beta))$$

almost everywhere, and the rent of type β is

$$U(\beta) = \int_{\beta}^{\bar{\beta}} \psi'(\tilde{\beta} - c(\tilde{\beta})) d\tilde{\beta}.$$

Necessary and sufficient second-order conditions are

$$\dot{c}(\beta) \geq 0$$

(see Roger Guesnerie and Laffont, 1984). Since $\dot{U}(\beta) < 0$, the firm's individual rationality constraint reduces to

$$U(\bar{\beta}) \geq 0.$$

APPENDIX 2

In addition to the bypass regime, we have potentially as many regimes as combinations of binding constraints among (8)–(11). However, the following lemmas cut down the number of cases to five.

LEMMA 1: *If the two types of consumers are offered two different contracts, the two incentive constraints cannot be simultaneously binding.*

PROOF:

Suppose the contrary. If the type-1 constraint is binding, $T_2 - T_1 = S(q_2) - S(q_1)$. If the type-2 incentive constraint is also binding, we have $\theta(S(q_2) - S(q_1)) = S(q_2) - S(q_1)$, a contradiction unless $q_1 = q_2$; but then $T_1 = T_2$, contradicting the fact that we have two distinct contracts.

LEMMA 2: *If the type i ($i = 1, 2$) incentive constraint is not binding, then the type- j individual rationality constraint is binding.*

PROOF:

Reduce T_j if the type- j individual rationality constraint is not binding.

LEMMA 3: *A pooling contract can never be optimal.*

PROOF:

Note first that in a pooling contract (q, T) consumers' incentive constraints are automatically satisfied. (i) If $p_2 = \theta S'(q) > (\beta - e)(1 + \lambda)$, increase q_2 by ε and the transfer by $dT_2 = \theta S'(q)\varepsilon$ so that type-2 consumers remain indifferent. Since at (q, T) the marginal rate of substitution between q

and T_1 is higher for type-2 consumers, this new allocation is incentive compatible. This raises welfare by $\alpha_2[\theta S'(q) - (1 + \lambda)(\beta - e)]\varepsilon$. (ii) If $p_2 \leq (\beta - e)(1 + \lambda)$, $p_1 < (\beta - e)(1 + \lambda)$. Decrease q_1 by ε adjusting T_1 so that type-1 consumers remain on the same indifference curve [i.e., $dT_1 = -S'(q)\varepsilon$]. Then, the total welfare change is $-(1 + \lambda)\alpha_1 S'(q)\varepsilon + (1 + \lambda)(\beta - e)\varepsilon$ which is positive by assumption, a contradiction.

Combining Lemmas 1, 2, and 3 we have only the five regimes described in the text in addition to the bypass regime.

APPENDIX 3

PROOF OF PROPOSITION 1:

The relevant transfers are deduced from the relevant binding constraints. For regime 1,

$$\begin{aligned} & \max\{\alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\ & \quad - (1 + \lambda)c(\alpha_1 q_1 + \alpha_2 q_2) \\ & \quad + \lambda[\alpha_1 S(q_1) + \alpha_2(\theta S(q_2) \\ & \quad \quad - \theta S(q_1) + S(q_1))]\} \end{aligned}$$

yields

$$\begin{aligned} \frac{p_1 - c}{p_1} &= \frac{\lambda}{1 + \lambda} \left(\frac{\alpha_2}{\alpha_1} \right) (\theta - 1) \\ p_2 &= c \end{aligned}$$

[letting $p_1 \equiv S'(q_1)$ and $p_2 \equiv \theta S'(q_2)$]. Note that

$$\frac{dp_1}{dc} > 0.$$

For regime 2,

$$\begin{aligned} & \max\{\alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\ & \quad - (1 + \lambda)c(\alpha_1 q_1 + \alpha_2 q_2) \\ & \quad + \lambda(\alpha_1 S(q_1) + \alpha_2(\theta S(q_2) - S_2^*))\} \end{aligned}$$

with the constraint

$$(\theta - 1)S(q_1) = S_2^*$$

yields

$$p_2 = c$$

$$q_1 = \bar{q} \equiv S^{-1} \left(\frac{S_2^*}{\theta - 1} \right)$$

implying

$$\frac{dp_1}{dc} = 0.$$

For regime 3,

$$\begin{aligned} & \max\{\alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\ & \quad - (1 + \lambda)c(\alpha_1 q_1 + \alpha_2 q_2) \\ & \quad + \lambda(\alpha_1 S(q_1) + \alpha_2(\theta S(q_2) - S_2^*))\} \end{aligned}$$

yields

$$p_1 = p_2 = c.$$

For regime 4,

$$\begin{aligned} & \max\{\alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\ & \quad - (1 + \lambda)c(\alpha_1 q_1 + \alpha_2 q_2) \\ & \quad + \lambda(\alpha_1 S(q_1) + \alpha_2(\theta S(q_2) - S_2^*))\} \end{aligned}$$

with the constraint

$$\theta S(q_2) - S(q_2) = S_2^*$$

yields

$$p_1 = c$$

$$q_2 = \bar{q} = S^{-1} \left(\frac{S_2^*}{\theta - 1} \right)$$

implying

$$\frac{dp_2}{dc} = 0.$$

For regime 5,

$$\begin{aligned} \max\{ & \alpha_1 S(q_1) + \alpha_2 \theta S(q_2) \\ & - (1 + \lambda) c (\alpha_1 q_1 + \alpha_2 q_2) \\ & + \lambda (\alpha_1 (S(q_1) - S(q_2)) + \theta S(q_2) - S_2^*) \\ & + \alpha_2 (\theta S(q_2) - S_2^*) \} \end{aligned}$$

yields

$$p_1 = c$$

$$\frac{p_2 - c}{p_2} = -\frac{\lambda}{1 + \lambda} \left(\frac{\alpha_1}{\alpha_2} \right) \left(\frac{\theta - 1}{\theta} \right)$$

implying

$$\frac{dp_2}{dc} > 0.$$

For regime 6,

$$\begin{aligned} \max\{ & \alpha_1 S(q_1) + \alpha_2 S_2^* - (1 + \lambda) c \alpha_1 q_1 \\ & + \lambda \alpha_1 S(q_1) \} \end{aligned}$$

yields

$$p_1 = c.$$

Note that since individual prices p_1 and p_2 are nondecreasing in all regimes, individual quantities q_1 and q_2 (and also aggregate quantities Q) are nonincreasing in c .

APPENDIX 4

PROOF OF PROPOSITION 2:

(i) Let Q^i denote aggregate production in regime i . Note that $Q^i > Q^6$ for $i = 1, \dots, 5$ and for any marginal cost c . From Appendix 3, under bypass, $p_1 = c$ and $q_2 = 0$. For $i = 3, 4$, or 5 , $p_1 = c$ so that q_1 is the same as in regime 6; but since $q_2 > 0$ in any regime i , $i \leq 5$, $Q^i > Q^6$. For $i = 1$ or 2 , IC_2 and IR_1 are strictly binding, and no bypass occurs. Hence, $(\theta - 1)S(q_1^1) \geq S_2^*$. In regime 6, IR_1 is strictly binding, and bypass occurs. Hence, $(\theta - 1)S(q_1^6) \leq S_2^*$. This implies that $Q^i > Q^6$ for $i = 1$ or 2 .

From the envelope theorem, for any i , $(d/dc)\tilde{V}_1^i(c) = -(1 + \lambda)Q^i$. Therefore,

$$\begin{aligned} \frac{d}{dc} [\tilde{V}^i(c) - \tilde{V}^6(c)] \\ = -(1 + \lambda)(Q^i - Q^6) < 0 \quad i = 1, \dots, 5. \end{aligned}$$

Consequently, there exists $c^* \in [0, +\infty)$ such that regime 6 corresponds to c levels above c^* . Since c is a nonincreasing function of β , from (21), there exists an interval $[\beta^*, \bar{\beta}]$ (possibly degenerated) in which bypass occurs.

Note that, at β^* , there is a discontinuity in total production, since in regime 6,

$$S'(q_1) = (\beta^* - e)$$

$$\psi'(e) = \alpha_1 q_1 - \frac{\lambda}{1 + \lambda} \left(\frac{F(\beta^*)}{f(\beta^*)} \right) \psi''(e)$$

$$Q^6 = \alpha_1 q_1$$

and in regime 5,

$$S'(q_1) = (\beta^* - e)$$

$$\theta S'(q_2) = \frac{(\beta^* - e)}{1 + \frac{\lambda}{1 + \lambda} \left(\frac{\alpha_1}{\alpha_2} \right) \left(\frac{\theta - 1}{\theta} \right)}$$

$$\begin{aligned} \psi'(e) &= \alpha_1 q_1 + \alpha_2 q_2 \\ &\quad - \frac{\lambda}{1 + \lambda} \left(\frac{F(\beta^*)}{f(\beta^*)} \right) \psi''(e). \end{aligned}$$

The discontinuity in Q and e translates into a discontinuity in marginal cost at β^* .

(ii) By the change of variables $S(q_1) \equiv s_1$ and $S(q_2) \equiv s_2$, the constraints define a convex set in the space of control variables. Furthermore, because $\psi'' \geq 0$, the function of the control variable c , $-\psi(\beta - c)$, is concave, and the objective function is concave in control and state variables. Therefore, the Pontryagin conditions are sufficient and yield continuous controls on $[\beta, \beta^*]$ and $[\beta^*, \bar{\beta}]$; see theorem 5 in Atle Seierstad and Knut Sydsæter (1987 p. 28).

By ordering the regimes according to 1, 2, 3, 4, 5, we exhibit a continuous solution that satisfies the Pontryagin conditions and is therefore a solution.

APPENDIX 5

Suppose that we have regime i ($i \leq 5$) at the left of β^* . The second derivative of the objective function with respect to β^* is [using $d\tilde{V}^i/dc = -(1+\lambda)Q^i$]

$$\begin{aligned} \text{(A4)} \quad & -(1+\lambda)Q^i \frac{dc^i}{d\beta} - (1+\lambda) \\ & \times \psi'(\beta^* - c^i(\beta^*)) \left(1 - \frac{dc^i}{d\beta}\right) \\ & + \lambda \psi'(\beta^* - c^i(\beta^*)) \\ & + (1+\lambda)Q^6 \frac{dc^6}{d\beta} + (1+\lambda) \\ & \psi'(\beta^* - c^6(\beta^*)) \left(1 - \frac{dc^6}{d\beta}\right) \\ & - \lambda \psi'(\beta^* - c^6(\beta^*)). \end{aligned}$$

Using the fact that

$$\begin{aligned} & \psi'(\beta^* - c^i(\beta^*)) \\ & = Q^i - \frac{\lambda}{1+\lambda} \left(\frac{F(\beta^*)}{f(\beta^*)} \right) \psi''(\beta^* - c^i(\beta^*)) \end{aligned}$$

(A4) becomes

$$\begin{aligned} \text{(A5)} \quad & -[\psi'(\beta^* - c^i(\beta^*)) \\ & - \psi'(\beta^* - c^6(\beta^*))] \\ & - \lambda \frac{F(\beta^*)}{f(\beta^*)} \left[\psi''(\beta^* - c^i(\beta^*)) \frac{dc^i}{d\beta} \right. \\ & \left. - \psi''(\beta^* - c^6(\beta^*)) \frac{dc^6}{d\beta} \right]. \end{aligned}$$

Since $\beta^* - c^i(\beta^*) > \beta^* - c^6(\beta^*)$ for all i , (A5) is negative for $\lambda = 0$. The objective function is therefore concave in β^* in a neighborhood of $\lambda = 0$.

For λ large, the result may become ambiguous in some cases. For example, if ψ''' is constant, the second part of (A5) is also negative in regimes 3, 4, and 5. In regime 2, the sign of the second term is ambiguous because $|dQ/dc|$ may be smaller in regime 2 than in regime 6.

APPENDIX 6

It can be checked that each of the regimes is relevant for some values of the parameters. Of particular interest for our analysis is the possibility of existence of regime 5. To check this, let us construct economies for which regime 5 is optimal among regimes $i \in \{1, \dots, 5\}$ for all $\beta \in [\beta, \bar{\beta}]$ (and therefore is globally optimal for small β 's). Suppose that

$$S(q) = q^{1-(1/\varepsilon)} \left/ \left(1 - \frac{1}{\varepsilon}\right) \right.$$

where $\varepsilon > 1$. That is, the demand functions have constant elasticity ε . Consider a sequence of economies indexed by θ in which θ tends to 1 (the consumers become more and more alike). The bypass technology has cost $f(\theta) + d(\theta)q$, where $f(\cdot)$ and $d(\cdot)$ are to be determined. All other data are fixed. Straightforward computations show that

$$S_2^* - S_1^* = (d(\theta))^{1-\varepsilon} (\theta^\varepsilon - 1) \left/ \left(1 - \frac{1}{\varepsilon}\right) \right.$$

and

$$S_2^* = (d(\theta))^{1-\varepsilon} \theta^\varepsilon \left/ \left(1 - \frac{1}{\varepsilon}\right) \right. - f(\theta).$$

Now choose $d(\theta)$ converging to 0 sufficiently fast with θ so that $S_2^* - S_1^* \rightarrow +\infty$, and choose $f(\theta)$ so as to keep S_2^* constant. The analysis in the text is unchanged, as S_2^* is constant along the sequence and S_1^* is negative. However, $p_1(\theta)$ and $p_2(\theta)$ converge to the marginal cost $(\beta - e)$ in all

regimes. Because the difference $\theta S(q_2) - S(q_1)$ converges to 0 when θ converges to 1 (as long as the marginal cost $\beta - e$ does not converge to 0, which is guaranteed if β is not too small and ψ is sufficiently steep), in the limit $S(q_1) = S_2^* > 0$. Therefore, the type-1 customers' IR constraint is not binding, which indicates that regime 5 is obtained in the limit [if bypass is prevented, which will be the case for an appropriate $\psi(\cdot)$ function].

REFERENCES

- Baron, David and Myerson, Roger, "Regulating a Monopolist with Unknown Costs," *Econometrica*, July 1982, 50, 911-30.
- Baumol, William, Panzar, John and Willig, Robert, *Contestable Markets and the Theory of Industry Structure*, New York: Harcourt Brace Jovanovich, 1982.
- Caillaud, Bernard, "Regulation, Competition and Asymmetric Information," manuscript, 1985; (*Journal of Economic Theory*, forthcoming).
- Champsaur, Paul and Rochet, Jean-Charles, "Multiproduct Duopolists," *Econometrica*, May 1989, 57, 533-58.
- Demski, Joel, Sappington, David and Spiller, Pablo, "Managing Supplier Switching," *Rand Journal of Economics*, Spring 1987, 18, 77-97.
- Einhorn, Michael, "Optimality and Sustainability: Regulation and Intermodal Competition in Telecommunications," *Rand Journal of Economics*, Winter 1987, 18, 550-63.
- Guesnerie, Roger and Laffont, Jean-Jacques, "A Complete Solution to a Class of Principal-Agent Problems with an Application to the Control of a Self-Managed Firm," *Journal of Public Economics*, August 1984, 25, 329-69.
- Kahn, Alfred, *The Economics of Regulation: Principles and Institutions, Volume II: Institutional Issues*, New York: Wiley, 1971 (reprinted by MIT Press, Cambridge, MA, 1988).
- Laffont, Jean-Jacques and Tirole, Jean, "Using Cost Observation to Regulate Firms" *Journal of Political Economy*, June 1986, 94, 614-41.
- ____ and _____, "Comparative Statics of the Optimal Incentives Contract," *European Economic Review*, June 1987, 31, 901-26.
- ____ and _____, (1990a) "The Regulation of Multiproduct Firms, Part I: Theory," forthcoming, *Journal of Public Economics*, 1990.
- ____ and _____, (1990b) "The Regulation of Multiproduct Firms, Part II: Applications to Competitive Environments and Policy Analysis," forthcoming, *Journal of Public Economics*, 1990.
- ____ and _____, (1990c) "The Politics of Public Decision Making: Regulatory Institutions," *Journal of Law, Economics and Organization*, Spring 1990, 6, 1-32.
- Lewis, Tracy and Sappington, David, "Countervailing Incentives in Agency Problems," *Journal of Economic Theory*, December 1989, 49, 294-313.
- Maskin, Eric and Riley, John, "Monopoly with Incomplete Information," *Rand Journal of Economics*, Summer 1984, 15, 171-96.
- Mussa, Michael and Rosen, Sherwin, "Monopoly and Product Quality," *Journal of Economic Theory*, June 1978, 18, 301-17.
- Scharfstein, David, "The Disciplinary Role of Takeovers," *Review of Economic Studies*, April 1988, 55, 185-200.
- Seierstad, Atle and Sydsaeter, Knut, *Optimal Control Theory with Economic Applications*, Amsterdam: North-Holland, 1987.

Durable-Good Monopoly and Best-Price Provisions

By DAVID A. BUTZ*

Best-price provisions guarantee buyers that the prices they pay are the lowest available. If the seller subsequently cuts price, then each previous buyer is entitled to a refund. A durable-good monopolist who offers certain forms of these provisions can construct a consistent plan yielding the same profits as rental agreements and contracts with explicit quantity commitments. The provisions require special circumstances to be practical, but they are simple and effective and appear in a variety of economic settings. Three applications are discussed: international commodity agreements, markets for electric turbogenerators, and markets for financial claims. (JEL 610, 611, 612)

In a classic paper, Ronald Coase (1972) conjectures that a monopoly seller of an infinitely durable good cannot sell output at the static monopoly level. Once the initial quantity has been sold, more profits can be made by cutting price and increasing output. Profit opportunities end only after price falls to marginal cost. Without some restraints, the market is saturated with the competitive output "... in the twinkling of an eye" (Coase, 1972 p. 143).

Since monopolists do not routinely behave as competitors, either real-world conditions do not mirror those assumed in Coase's illustration or monopolists somehow commit not to behave in this manner. Nancy Stokey (1981), Jeremy Bulow (1982), and Charles Kahn (1986) show that if production capacity is limited or if marginal cost is increasing, then the monopolist's problem is less severe. Lawrence Ausubel and Raymond Deneckere (1987, 1989) demonstrate the existence of equilibria other than the one described by Coase and show that potential entry may actually mitigate the monopolist's problems. In short, circumstances may not be as bleak as in Coase's exposition. Nonetheless, the problem often remains in less severe form.

This suggests that monopolists commit not to behave competitively. Coase offers three possibilities. The monopolist can commit not to sell additional output; the good can be rented rather than sold; or the monopolist can agree to repurchase the good if ever a lower price is offered.

This paper offers an alternative similar to repurchase agreements.¹ The monopolist's problem can be resolved by employing best-price (BP) provisions.² These guarantee that the price to be paid or received is the best available. If better terms are subsequently negotiated in any related contract, then the monopolist must refund the difference between the original price and the new lower price. The outcome is the same as if the monopolist had repurchased the good at the original price and then resold it at the lower price. BP provisions and repurchase agreements therefore differ in only one respect: while repurchase agreements require the monopolist to reassume ownership of the good, BP provisions do not.

After providing a background (Section I), an explanation and example (Section II) illustrate the mechanics of BP provisions. A discrete-time model with demand uncertainty is outlined in Section III. The propo-

*Department of Economics, University of California at Los Angeles, Los Angeles, CA 90024-1477. I thank the participants in workshops at UCLA and USC, two anonymous referees, Bill Gale, John Riley, and especially Michael Waldman for helpful comments on earlier versions. All errors are my own.

¹The similarity between best-price provisions and repurchase agreements is also discussed by Ivan Png (1987).

²Cooper (1984) suggests best-price provisions to resolve the durable-good monopoly problem.

sitions in Section IV demonstrate the role BP provisions play in resolving the monopolist's problem with dynamic inconsistency. Section V discusses the results in a continuous-time framework. Section VI lists the provisions' advantages and disadvantages, Section VII provides some applications, and a conclusion follows.

I. Background on Best-Price Provisions

Although best-price provisions are pervasive in many economic contexts, their scope is typically restricted. Retailers, for example, often extend the provision only to the same brand name and model and limit it to specific time periods and geographic areas. "Three-party" BP provisions (or "meet-the-competition" clauses) guarantee the lowest price offered by *any* seller of the good; "two-party" versions apply only to the lowest price offered by the seller involved in the original transaction.

International commodity agreements have employed best-tariff terms, known as "most-favored-nation" (MFN) provisions, for over three centuries. By offering such provisions, a country promises each trading partner access to its domestic markets at tariff rates that are no higher than those offered by that country to any other trading partner. The consensus in the international trade literature is that MFN's assure nondiscrimination:

... the most-favored-nation clause conferred no privileges of any great importance, for the general rule was to treat all nations as equals. Most-favored-nation treatment then, meant not favored treatment, but merely a guarantee against being less favorably treated than other foreign nations. (Vernon Setser, 1937 p. 69)

Potential discrimination has also been cited to explain BP provisions in long-term contracts between natural-gas producers and pipelines (Edward Neuner, 1960; R. Glenn Hubbard and Robert Weiner, 1986).

Other authors (Frederic Scherer, 1980; David Grether and Charles Plott, 1984; Charles Holt and David Scheffman, 1987;

Stephen Salop, 1986; Thomas Cooper, 1986) argue that two-party BP provisions enhance tacit collusion. By committing the firm to pay rebates if it ever cuts price, the provisions reduce competition for new customers. Terrence Belton (1986) demonstrates how meet-the-competition provisions enhance collusion by committing each firm to match the prices of industry rivals.

Risk sharing is also a proposed motive for BP provisions in field markets for natural gas. Paul MacAvoy (1962) and Harry Broadman and W. David Montgomery (1983) argue that the provisions may result in a more efficient allocation of risk by shifting price uncertainty from the beneficiary of the clause to the benefactor.

Attention here focuses on the nondiscrimination motive for BP provisions. Both two- and three-party versions are modeled. Alternate hypotheses are raised again in the conclusion.

II. A Simple Explanation of Best-Price Provisions

The demand for ownership of a durable good is illustrated in Figure 1. The price, $P(X)$, is decreasing in the quantity X sold. If a monopoly seller produces at zero marginal cost, the solution might appear to involve selling M units at price $P(M)$ per unit. Coase's revelation is that this solution does not hold when future behavior is contemplated. Having sold M units, the monopolist can lower price to sell additional output. Knowing this, prospective buyers balk at paying $P(M)$. Unless the monopolist can commit not to cut price, no output can be sold at any price above marginal cost.³

Now consider the outcome when BP provisions are offered. Suppose the monopolist

³Coase's reasoning has been supported by several subsequent authors, including Stokey (1981), Bulow (1982), Eric W. Bond and Larry Samuelson (1984), Faruk Gul, Hugo Sonnenschein, and Robert Wilson (1986), and Kahn (1986). When marginal cost is increasing, price does not immediately fall to marginal cost, but in all cases, the monopolist may be unable to reap the full rewards that would be possible through commitment to a preannounced production plan.

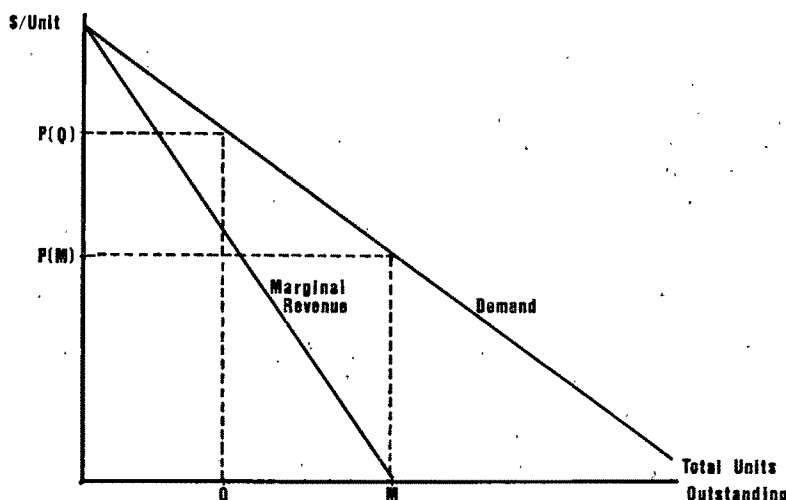


FIGURE 1. DEMAND FOR OWNERSHIP OF A DURABLE GOOD

begins by selling Q units at price $P(Q)$ per unit. If ΔQ additional units are sold, revenues increase by $P \times \Delta Q$, but the monopolist must rebate $Q \times \Delta P$ to previous customers. Output is set such that marginal rebates equal marginal revenue. This yields the standard monopoly outcome.

A. An Example

This result also holds when cost or demand is uncertain. A hypothetical event (say an art show) is funded in part through sales of a commemorative lithograph. There is only one seller of the prints, the event's sponsor, who produces at zero marginal cost and maximizes expected profits. Sales take place before ($t=0$) and after ($t=1$) the event. The sponsor can credibly commit to destroy the plate after the second period, so there are only two production decisions, q_0 and q_1 .

Potential buyers include 1,000 individuals who will pay up to \$100 for the print. An additional η individuals will pay up to \$60, where η is distributed uniformly over $[0, 1,000]$, and observed only after the event occurs (i.e., at time $t=1$). Figure 2 shows final demand for the print.

Up to 1,000 prints can be sold for \$100 at time $t=0$, but not if customers foresee the

possibility of a \$60 price one period later. The monopolist could commit not to produce more than 1,000 prints but would like the flexibility to sell additional amounts if η is high. The seller maximizes expected profits by committing not to sell prints at time $t=1$ unless $\eta \geq 667$. The 1,000 prints sold at time $t=0$ command a price which reflects the $1/3$ probability of a price cut in the following period.⁴

Suppose BP provisions are used instead. The seller sets $q_0 = 1,000$ and charges \$100 per print. At time $t=1$, the seller then weighs the revenues from selling η additional prints ($\$60 \times \eta$) against the rebates that would have to be paid to previous buyers ($\$40 \times 1,000$). The seller sets $q_1 = \eta$ if and only if $\eta \geq 667$. Otherwise $q_1 = 0$. Because BP provisions redistribute risk from initial buyers to the seller, the initial price and realized profits differ from the scenario in which quantity commitments are employed. Yet output levels and expected profits are the same.

This example is discussed in further detail in Section VII.

⁴If buyers are risk-neutral and have the same discount rate, δ , then the initial price equals $\phi_0 + \delta\{(2/3)(\$100) + (1/3)(\$60)\}$, where ϕ_0 is the rental value of the good in the initial period.

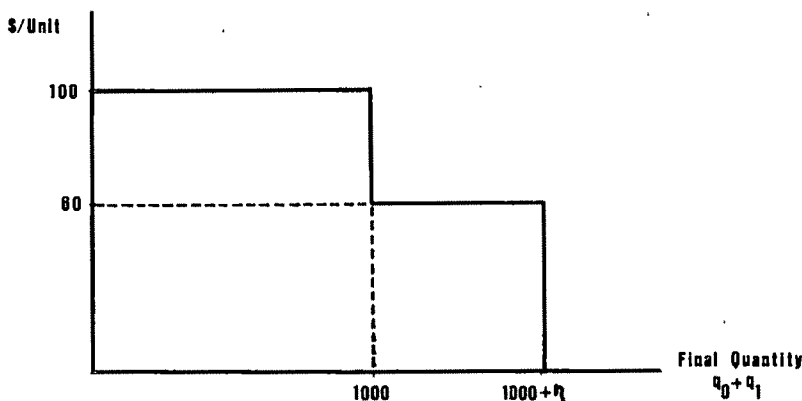


FIGURE 2. FINAL DEMAND FOR THE LITHOGRAPH IN THE EXAMPLE (see text)

III. The Model

At time $t = 0$ a (new) durable good is introduced. It never depreciates and is available initially only through a monopoly seller. Once purchased, it can be leased or resold through perfectly competitive secondary markets.

The total stock of the good at time t is Q_t , and $q_t = (Q_t - Q_{t-1})$ is the quantity sold by the monopolist at that time. The inverse demand for the good's rental services is $\phi_t = \phi_t(Q_t, \varepsilon_t)$, where ε_t is a random variable with probability density function $f_t(\varepsilon_t)$. Production costs are $\gamma_t = \gamma_t(q_t)$, where $\gamma'_t(q_t) \geq 0$ and $\gamma''_t(q_t) \geq 0$.⁵ There is a common discount rate, δ , and all agents know γ_t , f_t , and ϕ_t .

Output can be priced in a variety of ways. Assume that buyers are risk-neutral. Let β_s^t be the payment by a time- t buyer to the monopolist at time $s \geq t$. The monopolist's choice of pricing conventions must satisfy

$$(1) \quad \beta_t^t + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \beta_s^t \\ = \phi_t(Q_t, \varepsilon_t) + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s(Q_s, \varepsilon_s)$$

for all t . The first left-hand-side term, β_t^t , is

⁵Cost uncertainties could also be introduced.

the buyer's initial payment. Expected discounted payments in subsequent periods appear in the second expression and can be positive or negative. If the monopolist rents the good, then $\beta_s^t = \phi_s$ for all t and all $s \geq t$. With "simple" prices (unaccompanied by price guarantees), $\beta_t^t > 0$ and $\beta_s^t = 0$ for $s > t$.

The right side of equation (1) measures the expected value of the good's rental services. The equation therefore constrains the monopolist to choose a pricing scheme such that expected discounted payments equal the expected discounted value of the rental services provided. Three pricing mechanisms are outlined in this section; all satisfy this constraint.

A. Simple Prices

Without price guarantees, the price at time t is

$$(2) \quad P_t = \phi_t(Q_t, \varepsilon_t) \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s(Q_s, \varepsilon_s).$$

The firm's discounted cash flow from time t onward is given by

$$(3) \quad \Pi_t = \{P_t q_t - \gamma_t(q_t)\} \\ + \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{P_s q_s - \gamma_s(q_s)\}$$

Let $\{q_t^*, t = 0, 1, \dots\}$ be the contingent production plan maximizing $E_0 \Pi_0$. This plan is dynamically consistent if it also maximizes $E_t \Pi_t$ for all t . In general, $\{q_t^*, t = 0, 1, \dots\}$ is not consistent when simple prices are employed.

For reasons to be explained shortly, consider only cases where

$$(4) \quad \phi_t(Q_t^*, \varepsilon_t) \geq \phi_{t+1}(Q_{t+1}^*, \varepsilon_{t+1})$$

for all t . This assumption assures that prices do not rise over time.

B. Infinite-Duration, Two-Party Best-Price Provisions

Suppose BP guarantees extend forever but apply only to subsequent transactions with the monopolist. Formally, suppose the monopolist guarantees anyone with such a provision at time t that, for all s and all $k \leq s - t$,

$$(5) \quad \sum_{i=t}^s \beta_i^t \leq \sum_{j=t+k}^s \beta_j^{t+k}.$$

In short, the monopolist promises each time- t buyer that its cumulative payments will not exceed the payments made by any subsequent buyer.

Let $\rho_t = \beta_t^t$ be the monopolist's time- t price when BP provisions are offered. Then at time $t+1$, new buyers pay ρ_{t+1} and time- t buyers receive a rebate of $(\rho_{t+1} - \rho_t) = -\beta_{t+1}^t$. This brings the net cost of the good for time- t buyers to ρ_{t+1} . In the same fashion, anyone who has purchased a unit of the good through time $t+s-1$ receives a rebate of $(\rho_{t+s-1} - \rho_{t+s})$ at time $t+s$. The net price at time $t+s$ of one unit purchased at time t therefore equals ρ_{t+s} . All buyers, regardless of their vintage, pay this same "best" price.

The sale price at time t is determined by the following:

$$(6) \quad \rho_t - E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} (\rho_{s-1} - \rho_s) \\ = \phi_t(Q_t, \varepsilon_t) \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s(Q_s, \varepsilon_s).$$

Through algebraic manipulation, (6) implies that

$$(7) \quad \rho_t = (1 - \delta)^{-1} \phi_t = \sum_{s=0}^{\infty} \delta^s \phi_{t+s}.$$

Best-price provisions fully compensate buyers for any change in the value of the good. Hence, ρ_t is set as if inverse demand forever equals ϕ_t .

If the monopolist follows the plan $\{q_t^*, t = 0, 1, \dots\}$, then (4) and (7) imply that $\rho_{t-1} \geq \rho_t$ for all t . Thus, the rebate paid to each buyer at time t is always nonnegative.⁶

When the monopolist offers two-party BP provisions, discounted cash flow from time t onward is given by⁷

$$(8) \quad \Omega_t = \{\rho_t q_t - (\rho_{t-1} - \rho_t) Q_{t-1} - \gamma_t(q_t)\} \\ + \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{\rho_s q_s - (\rho_{s-1} - \rho_s) \\ \times Q_{s-1} - \gamma_s(q_s)\}.$$

C. One-Period Meet-the-Competition Provisions

Now consider the monopolist's problem if buyers are promised meet-the-competition (MTC) provisions extending for a single period. These provisions are frequently extended to buyers of such consumer durables as appliances and electronics. Under this arrangement, all buyers paying α_t for the good at time t receive a refund of $(\alpha_t - P_{t+1})$ at time $t+1$. The seller modeled here has a monopoly on primary sales, so all competition comes through the secondary market.⁸

⁶While the analysis that follows does not change if the assumption given by (4) is dropped, BP provisions are rarely observed in practice where prices are rising through time. In such cases, it would be necessary for the monopolist to collect "surcharges" whenever $\rho_{t-1} < \rho_t$.

⁷It is assumed that $\rho_{-1} = Q_{-1} = 0$.

⁸Since the model assumes monopoly, the term "meet the competition" may be somewhat confusing. Here atomistic buyers and sellers in the secondary market individually have no impact on price. They also have rational expectations regarding the monopolist's quan-

At time t , the monopolist's price, α_t , is determined by

$$(9) \quad \alpha_t - \delta E_t \{\alpha_t - P_{t+1}\} = \phi_t(Q_t, \varepsilon_t) + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s(Q_s, \varepsilon_s).$$

From (9) it follows that

$$(10) \quad \alpha_t = (1 - \delta)^{-1} \phi_t = \sum_{s=0}^{\infty} \delta^s \phi_t.$$

As before, price is set as if inverse rental demand forever equals ϕ_t . By (4) and (10), $(\alpha_t - P_{t+1}) \geq 0$, so buyers always receive nonnegative rebates.

With MTC provisions, the monopolist's discounted cash flow from time t onward is given by⁹

$$(11) \quad \Gamma_t = \{\alpha_t q_t - (\alpha_{t-1} - P_t) q_{t-1} - \gamma_t(q_t) + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{\alpha_s q_s - (\alpha_{s-1} - P_s) q_{s-1} - \gamma_s(q_s)\}.$$

Having described these various options for pricing output, attention now turns to their impact on the dynamic consistency of the monopolist's plans.

IV. The Impact of Best-Price and Meet-the-Competition Provisions

This section provides three results. First, for any given production plan, price guarantees can be extended from the outset without altering expected profits. Second, if the monopolist offers two-party BP provisions, then the production plan $\{q_t^*, t = 0, 1, \dots\}$ is dynamically consistent. In contrast, one-period MTC provisions mitigate but do not resolve the monopolist's problems with dy-

namic inconsistency. Finally, the pricing provisions themselves are dynamically consistent: once the monopolist adopts BP or MTC provisions, there is no gain from dropping them at a later date.

A. Expected Profits

For the moment, ignore problems with dynamic inconsistency. For any given production plan, the monopolist's expected revenue per unit, given in equations (1), (2), (6), and (9), is the same. Buyers pay only for the expected rental services they consume. They may initially pay more when offered BP guarantees, but only if they expect future rebates. Formally, it is shown in the Appendix that for any plan $\{q_t, t = 0, 1, \dots\}$,

$$E_0 \Pi_0 = E_0 \Omega_0 = E_0 \Gamma_0 = \{\phi_0 q_0 - \gamma_0(q_0) + E_0 \sum_{t=1}^{\infty} \delta^t \{\phi_t Q_t - \gamma_t(q_t)\}.$$

B. Consistency of Production Plans

Two-party BP provisions protect owners of the durable good from all future changes in its value. When choosing output, the monopolist therefore internalizes the change in value of all units sold previously. One-period MTC provisions compensate buyers only from the previous period, so the monopolist considers the welfare of only a fraction of former customers. Hence, MTC provisions mitigate but do not resolve the dynamic inconsistency problem. The first proposition, proved in the Appendix, states these results formally.

PROPOSITION 1. If the monopolist offers infinite-duration, two-party BP provisions to all buyers, then the production plan $\{q_t^*, t = 0, 1, \dots\}$ is dynamically consistent. However, this plan is not in general consistent when one-period MTC provisions are employed.

One-period MTC provisions are useful but not completely effective. Unlike two-party BP provisions, however, they require only a one-time rebate and may be less expensive to administer. One-period MTC provisions

tity decisions, as well as future prices and rental values. Whenever price exceeds the expected value of the good's discounted flow of rental services, anyone holding the good attempts to sell it. Whenever price falls below the value of these services, the quantity demanded exceeds the total stock of the good outstanding.⁹

⁹It is assumed that $\alpha_{-1} = Q_{-1} = 0$.

may also give the monopolist some discretion to discriminate intertemporally when secondary markets are imperfect (see Section VII). Finally, MTC provisions can be coupled with other commitments to enhance their effectiveness.

One question still remains: does the monopolist at any point have an incentive to switch pricing plans? In other words, is the adoption of best-price provisions dynamically consistent?

C. Consistency of Pricing Plans

The specifications of Ω_t and Γ_t assume that, once the monopolist adopts BP or MTC provisions, they are also offered to all subsequent buyers. Yet by changing pricing policies, the monopolist can alter the timing—and perhaps the magnitude—of the rebates paid to previous customers. Could such a switch increase expected profits? The second proposition, proved formally in the Appendix, provides the answer.

PROPOSITION 2: *Suppose the monopolist has offered BP or MTC provisions through time $t-1$. Then, at time t the monopolist cannot increase expected discounted future cash flow by changing pricing policies.*

Price guarantees index the payments for units purchased in one period to the payments for units purchased subsequently. The expected discounted value of the payments for units purchased at time t must equal the value of the rental services provided [by eq. (1)]. The monopolist can alter the timing of rebates, but not their (discounted) magnitude. Hence, the monopolist has nothing to gain by switching to a different pricing policy at any future date. With two-party BP provisions, the monopolist's output decision and pricing policy are both dynamically consistent. With one-period MTC provisions, the monopolist's pricing policy is dynamically consistent, but the output decision is not.

V. The Length of Period

Though the model is most naturally expounded in discrete time, the results may be

sensitive to the length of the period (see Stokey, 1981; Kahn, 1986). In discrete time, the monopolist has a slight ability to commit; once output has been chosen in one period, it cannot be changed until the next. As the length of the period decreases, so does the monopolist's ability to commit in this manner.

With one-period MTC provisions, the monopolist internalizes the change in value of only those goods sold in the previous period. If the period is shortened, then the monopolist's production decisions reflect the impact on a smaller fraction of all units sold previously, and the effectiveness of these provisions declines. Yet because the monopolist can respond by extending the duration of MTC guarantees, period length is not important.

Infinite-duration, two-party BP provisions commit the monopolist to internalize the market value of all units sold previously. Altering period length does not diminish the effectiveness of this commitment in any way.

VI. The Relative Merits of Best-Price Provisions

The problem of dynamic inconsistency can be addressed in a variety of ways, including not only best-price guarantees, but also rental arrangements, repurchase provisions, and output commitments. International commodity agreements (see Sections II and VIII) routinely extend best-tariff guarantees, while artists seem to prefer "limited editions" of their work. Repurchase agreements and durable-good rentals are possible but rarely observed. This section addresses the following question: what factors lead to the use of best-price guarantees in some circumstances but not others?

Some advantages of BP guarantees are immediately apparent. They often redistribute risk efficiently and grant the monopolist maximum flexibility to choose future output. If prices are publicly observable, then monitoring and enforcement costs are low. Unlike rental or repurchase agreements, BP provisions do not obligate the monopolist to reassume ownership of the good. Hence, they are more attractive when the good has some *ex post* specificity.

Although the model assumes rational and homogeneous expectations, best-price provisions may be especially attractive when expectations differ. As an illustration, consider the following numerical example. As before, suppose the seller knows that η is uniformly distributed over the interval $[0, 1,000]$, but now assume that buyers (mistakenly) believe that η is distributed uniformly over $[0, 2,000]$. If the seller employs quantity commitments, the promise is the same as before: $q_1 = \eta$ if $\eta \geq 667$ and $q_1 = 0$ otherwise. But now the parties cannot agree on a time-0 price. Buyers believe there is a $2/3$ chance of a price cut at time $t = 1$, even though the true probability is only $1/3$. Hence, they are not willing to pay what the prints are worth. Rather than accept a low price, the seller prefers to defer sales until time $t = 1$.

With best-price provisions, buyers pay \$100 for the print at time $t = 0$ regardless of their expectations, and they receive a rebate of \$40 at time $t = 1$ if and only if the monopolist sells additional output. Although buyers think a price cut is likely, this affects neither their willingness to pay nor the seller's output decisions.

If these expectations are reversed, then the seller prefers quantity commitments to BP provisions. However, if buyers know they are not well-informed, then BP provisions are preferred whether buyers are more or less optimistic than the seller. To illustrate, suppose a retail seller of consumer appliances knows what prices will be in the next period, but customers do not. Even if the retailer has no market power, customers are reluctant to pay full price, since they might miss out on a sale. Best-price provisions assure them that they can buy their appliances now and still take advantage of a lower price offered in the next period.¹⁰

Since best-price provisions can be restricted to specific geographic areas, brand names, or time periods, they permit limited pursuit of both intratemporal and intertem-

poral price discrimination.¹¹ This may help to explain why MTC provisions are used even though they do not fully address problems with dynamic inconsistency.

There are also clear disadvantages to best-price provisions. In practice, they are rarely employed when prices are rising over time. If $\phi_t(Q_t^*, \varepsilon_t) < \phi_{t+1}(Q_{t+1}^*, \varepsilon_{t+1})$, then $(\rho_t - \rho_{t+1})$ and $(\alpha_t - P_{t+1})$ are negative. Instead of a rebate, customers pay a "surcharge." In addition to the obvious reluctance of customers to pay such a levy, it may be difficult for the monopolist to track down former customers.

If prices are falling monotonically over time, then two-party versions of the clause require recurring refunds. Best-price provisions have the greatest appeal, therefore, when subsequent payments are either unlikely or inexpensive to distribute.

Perhaps the greatest complications arise with heterogeneous products. Best-price provisions prevent only intertemporal price discrimination. If the monopolist can offer subsequent customers higher quality, better warranties, free delivery, or other perks, then the protection offered by best-price provisions may be worth very little.

This problem can be addressed by adjusting for product differences and by promising most-favorable treatment along other economically relevant dimensions of the contract. Long-term contracts between natural-gas pipelines and producers often promise producers the best quality-adjusted price and contain prorationing provisions to prevent quantity discrimination. Nonetheless, heterogeneity increases the cost of using BP provisions.

Summarizing, several conditions must hold before best-price provisions can be employed successfully. Prices must be publicly observable and must not be rising over time; refunds must be either infrequent or inexpensive to distribute; and the product must be roughly homogeneous. The provisions are relatively more advantageous when buyers are more risk-averse or less well-informed than the monopolist or when the good has

¹⁰ Most retailers of consumer appliances probably do not wield significant market power. Hence, asymmetric information about future demand may provide a better explanation for BP provisions in this setting.

¹¹ Intertemporal price discrimination would work only if secondary markets are imperfect.

some *ex post* specificity. Relative to quantity commitments, best-price provisions are especially attractive when future contingencies are difficult to outline *ex ante* and to verify *ex post*.

VII. Applications

My model employs highly restrictive assumptions, including rational and homogeneous expectations, risk neutrality, perfect secondary markets, infinite durability, pure monopoly, and a common discount rate. These can all be relaxed. In the numerical example, best-price provisions work even when expectations are irrational and regardless of buyers' risk preferences. The seller has a monopoly over a certain type of print, but this hardly constitutes a "pure" monopoly. The illustration assumes no secondary markets and never mentions individual rates of time discount or infinite durability.

Only three conditions appear to be necessary: asset durability, seller market power, and forward-looking expectations. Under these circumstances, the seller must assure buyers that the value of their assets will not be diluted through excessively high output levels in future periods. This section discusses the role played by best-price and nondiscrimination clauses in settings where these conditions arise.

A. International Commodity Agreements

A country controls access to its domestic market for some good and licenses foreign trading partners to sell output there. These licenses are long-lived and paid for through reciprocal concessions and tariff revenues. For example, the United States might license Japanese car sales in the United States in exchange for a license to market wheat in Japan. The agreement might specify a \$500 tariff on each car and \$1 tariff on each bushel of wheat.

The value of these licenses depends on the number and type of licenses granted to other trading partners. After negotiating the treaty described above, suppose the United States agrees to a \$100 tariff on

West German cars. This agreement significantly reduces the value of Japan's license to sell cars here. If Japan had contemplated this scenario when negotiating their agreement, they would have demanded compensation.

In practice, compensation comes through most-favored-nation (MFN) provisions. An MFN provision would assure Japan that its tariff will be reduced to \$100 once the U.S.-German accord is signed.

These provisions are not problem-free. First, they promise only nondiscriminatory tariffs and can be circumvented through quotas and other nontariff trade barriers. Second, imports are rarely homogeneous, so countries can discriminate through their product classifications (e.g., the Suzuki Samurai jeep could be classified as a recreational vehicle rather than an automobile). Third, countries may wish to discriminate in favor of some trading partners (e.g., allies or lesser-developed countries), and MFN provisions offer only limited opportunities for doing so.

Finally, disputes often arise over the value of reciprocal tariff concessions. If the U.S.-German accord allows duty-free sales of U.S. wheat in West Germany, should Japan be obligated to offer the same terms, or should its tariff be reduced to \$100 regardless of the German concessions?

Each of these problems can be addressed by adding language to the treaty. Nontariff trade barriers can be prohibited, product definitions can be more precise, and exceptions can be made for preferred trading partners. Because the value of reciprocal concessions is difficult to measure, MFN provisions typically apply unconditionally. In our example, the damages to Japanese car makers are the same regardless of the concessions granted to Germany. Hence, it is far simpler to ignore them when determining tariff adjustments.¹²

¹²Unconditional MFN treatment results in "spillover effects." Japan effectively free-rides on the U.S. and German efforts to liberalize trade. Spillover effects have been a primary reason why countries have moved toward multilateral tariff bargaining.

B. The Market for Electric Turbogenerators

Perhaps the most infamous application of best-price provisions is in the market for electric turbogenerators. Prior to 1963, this market was characterized by elaborate collusive efforts, as well as chronic price wars. In 1963, General Electric introduced a "price protection plan," and Westinghouse, its largest rival, soon followed suit. Each plan guaranteed that if the firm gave a discount on any new turbogenerator order, the same discount would be offered retroactively on all orders taken within the previous six months. At the same time, the firms adopted simplified booking procedures to standardize pricing and opened their records for public scrutiny.

These policies made it unprofitable to cut prices, since any discounts to one customer involved rebates to all others. After some initial rivalry, the firms established price stability, and there was not one price cut by either firm until the plans were terminated by government order in 1977.

Although tacit collusion is an obvious explanation for these provisions (see Section II), there is a complementary interpretation: potential buyers, aware of recurring price wars, did not find cartel behavior credible. Whenever possible, they postponed orders, hoping to take advantage of the next round of discounts. Best-price provisions assured buyers that they could place their orders and still take advantage of subsequent reductions. At the same time, they committed the firms not to cut price.

In the conventional explanation, best-price provisions enhance cooperation between firms at each point in time. Here the provisions enable a single cartel to collude with itself across time. Best-price provisions could be commitments to industry rivals, yet they could also serve as commitments to buyers. The two explanations are perfectly compatible and together lead to a richer model. Even if the firms were able to collude, they would still need a dynamically consistent plan; and dynamic consistency would not have been an issue if the firms were competitors.

C. Discrimination Between Stockholders

Suppose a corporation is financed solely through the sale of shares of common stock. Let V_t be the aggregate value of these shares at time t . Then if N_t is the number of claims outstanding, the price per share is

$$(12) \quad P_t = \frac{V_t}{N_t}.$$

Now suppose at this time that $n_t > 0$ additional shares are sold in a financial transaction.¹³ The firm charges p_t per share, and the proceeds, $p_t n_t$, are retained by the firm. The value of the original N_t shares is affected in two ways. First, each share's proportionate claim on the firm falls. Instead of owning the fraction $1/N_t$, each share now commands only $1/(N_t + n_t)$. Second, the proceeds are retained, so the aggregate value of the firm rises to V'_t ,¹⁴ where

$$(13) \quad V'_t = V_t + p_t n_t.$$

The two effects work in opposite directions: the original shares each command a smaller fraction of a larger pie. The new price, P'_t , is

$$(14) \quad P'_t = \frac{V'_t}{N_t + n_t} = \frac{V_t + p_t n_t}{N_t + n_t}.$$

If $p_t = P_t$, then from equations (12) and (14) it follows that $P'_t = P_t$. In other words,

¹³By assuming that the exchange is purely financial, I rule out the possibility that the stock is offered in exchange for services rendered. Thus, I rule out cases in which stock is offered to employees, management, or board members as part of an overall compensation package. I also rule out cases involving contests for corporate control, since such contests involve not only the exchange of financial assets, but also the manner in which real resources are allocated.

¹⁴This discussion merely illustrates the role of discrimination. In a formal proof, it would be assumed that the proceeds of the stock sale, $p_t n_t$, are invested in assets that can be bought and sold by individual investors on the same terms available to the corporation. When coupled with other assumptions adopted in propositions on the irrelevance of financial policy, this assures that the value of the firm changes by exactly $p_t n_t$.

shareholders are indifferent to the sale (repurchase) whenever the terms are nondiscriminatory. The revenues retained from the sale of stock increase the firm's value by just enough to compensate for the reduction in each share's proportionate claim. If $p_i < P_i$, then $P'_i < P_i$. If the sale is discriminatory, shareholders are unambiguously worse off.

Suppose the firm does not promise to refrain from discriminatory transactions. If investors anticipate discrimination, then stock prices and the value of the firm fall before the discrimination occurs.

This last conclusion appears to be at odds with propositions showing that financial policy is irrelevant. Yet most irrelevance propositions, including Franco Modigliani and Merton Miller's (1958), do not address the dynamics of financial policy. The proposition closest in spirit to this discussion is outlined by Eugene Fama (1978). Fama shows that financial policy has no effect on the firm's value even when it makes no commitments regarding future sales of financial claims. However, Fama assumes that all transactions take place at market prices and thereby rules out discrimination. Fama's result should therefore be amended to say that financial policy is irrelevant even if the only commitment is to refrain from discriminatory financial transactions.

Discrimination becomes especially problematic when shareholders are heterogeneous. The firm would like to commit not to enter into discriminatory transactions that redistribute wealth from one class of shareholders to another. By doing so, it lowers its cost of capital. Yet discrimination may be necessary to compensate shareholders engaged in costly but value-enhancing contests for corporate control (Sanford Grossman and Oliver Hart, 1980). Controversies surrounding targeted share repurchases illustrate the difficulties of pursuing both objectives simultaneously.

VIII. Conclusions

Although the Coase conjecture is expounded using very specific assumptions, the problem of dynamic inconsistency can arise whenever a seller (or buyer) of a durable

asset possesses market power. Best-price provisions represent one mechanism for addressing this problem. They are practical only under special circumstances but are remarkably simple and effective.

When extended to all buyers, best-price provisions guarantee equal treatment. Equal-treatment guarantees appear in a variety of settings and are often justified on purely normative grounds. Through the model and applications, this paper outlines a positive analysis of these provisions.

While it is individually rational for buyers to accept best-price provisions, in equilibrium it appears that consumers as a whole suffer rather than benefit. Yet there are various reasons why the reader should not draw sweeping policy conclusions from this study. First, suppose a durable-good monopolist—or monopolistic competitor—has large fixed costs and low marginal cost. Then, without some means to raise price above marginal cost, revenues do not cover costs, and the monopolist produces nothing. Second, even when best-price provisions hurt consumers, they may be less pernicious than other alternatives the monopolist may employ to commit to the monopoly output. Third, in some settings, including international commodity treaties and financial contracts, best-price provisions and equal-treatment guarantees are widely credited with raising consumer welfare.

Problems with dynamic inconsistency could provide a motive for nondiscrimination provisions in other contexts, as well. Suppose a firm requires its employees to invest in specific and long-lived human capital. The return on these investments depends upon both the wage and the number of hours worked. Once the investments have been sunk, the firm has monopsony power, and workers face the danger that their quasi-rents will be expropriated. The firm can commit to wage rates and employment levels, but this hampers flexibility.

Instead the firm can promise most favorable treatment. Each worker is guaranteed wages and employment conditions at least as favorable as those offered to all other workers. If workers are heterogeneous, then those with more specific investments could

be guaranteed a higher wage and first priority in the allocation of hours.

While this explanation is not implausible, risk sharing, tacit collusion, and information asymmetries have been forwarded as competing explanations, and normative considerations may also play a role. If the model is to be extended to settings such as these, tests may be constructed to evaluate these competing hypotheses empirically.

APPENDIX

PROOF THAT $E_0\Pi_0 = E_0\Omega_0 = I_0\Gamma_0$:

Substitute equation (2) into (3) and take expected values:

$$\begin{aligned} (A1) \quad E_0\Pi_0 &= \left\{ \left[\phi_0 + E_0 \sum_{s=1}^{\infty} \delta^s \phi_s \right] q_0 - \gamma_0(q_0) \right\} \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \left\{ \left[\phi_t + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s \right] q_t - \gamma_t(q_t) \right\} \\ &= \{ \phi_0 q_0 - \gamma_0(q_0) \} \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \{ \phi_t Q_t - \gamma_t(q_t) \}. \end{aligned}$$

Substitute equation (7) into (8) and take expected values:

$$\begin{aligned} (A2) \quad E_0\Omega_0 &= \{ (1-\delta)^{-1} \phi_0 q_0 - \gamma_0(q_0) \} \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \{ (1-\delta)^{-1} \\ &\times (\phi_t Q_t - \phi_{t-1} Q_{t-1}) - \gamma_t(q_t) \} \\ &= \{ \phi_0 q_0 - \gamma_0(q_0) \} \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \{ \phi_t Q_t - \gamma_t(q_t) \}. \end{aligned}$$

After taking expected values, equation (11) implies

$$\begin{aligned} (A3) \quad E_0\Gamma_0 &= \{ \alpha_0 q_0 - \gamma_0(q_0) \} - \delta(\alpha_0 - E_0 P_1) q_0 \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \{ \alpha_t q_t - \delta(\alpha_t - P_{t+1}) q_t - \gamma_t(q_t) \}. \end{aligned}$$

By equations (2) and (9),

$$\begin{aligned} (A4) \quad \alpha_t - \delta(\alpha_t - E_t P_{t+1}) \\ = \phi_t + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \phi_s. \end{aligned}$$

Substitution of (A4) into (A3) yields

$$\begin{aligned} (A5) \quad E_0\Gamma_0 &= \{ \phi_0 q_0 - \gamma_0(q_0) \} \\ &+ E_0 \sum_{t=1}^{\infty} \delta^t \{ \phi_t Q_t - \gamma_t(q_t) \}. \end{aligned}$$

Comparing (A1), (A2), and (A5), it follows that $E_0\Pi_0 = E_0\Omega_0 = E_0\Gamma_0$.

To facilitate exposition, the proof of Proposition 2 precedes that of Proposition 1.

PROOF OF PROPOSITION 2:

Suppose first that the monopolist has offered MTC provisions through time $t-1$. The monopolist's cash flow from time t onward is given by F_t , where

$$\begin{aligned} (A6) \quad F_t &= \{ \beta_t^t q_t - (\alpha_{t-1} - P_t) q_{t-1} - \gamma_t(q_t) \} \\ &+ \sum_{s=t+1}^{\infty} \delta^{(s-t)} \left\{ \left(\sum_{k=t}^s \beta_s^k q_k \right) - \gamma_s(q_s) \right\}. \end{aligned}$$

The monopolist chooses the β 's to maximize $E_t F_t$ subject to the constraint given by equation (1). After taking expected values in equation (A6), equation (1) can be substituted to give the monopolist's unconstrained objective function:

$$\begin{aligned} (A7) \quad E_t F_t &= \{ \phi_t q_t - (\alpha_{t-1} - P_t) q_{t-1} - \gamma_t(q_t) \} \\ &+ E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{ \phi_s (Q_s - Q_{t-1}) \\ &\quad - \gamma_s(q_s) \}. \end{aligned}$$

In its unconstrained form, the monopolist's objective function is independent of the pricing conventions adopted.

Now suppose the monopolist has offered BP provisions through time $t-1$. Cash flow

from time t onward is G_t , where

$$(A8) \quad G_t = \left\{ \beta_t^t q_t + \left(\sum_{k=0}^{t-1} \beta_t^k q_k \right) - \gamma_t(q_t) \right\} \\ + \sum_{s=t+1}^{\infty} \delta^{(s-t)} \left\{ \left(\sum_{k=0}^s \beta_s^k q_k \right) - \gamma_s(q_s) \right\}$$

and the β 's are determined by the monopolist's BP obligations. The monopolist chooses pricing conventions to maximize $E_t G_t$ subject to the constraint given by equation (1).

If the monopolist chooses to continue offering BP provisions, then $\beta_t^t = \rho_t$ and $\beta_s^t = (\rho_s - \rho_{s-1})$ for all $s \geq t+1$. These results, together with equations (1), (7), and (A8), imply that

$$(A9) \quad E_t G_t = E_t \Omega_t \\ = \{ \phi_t Q_t - \rho_{t-1} Q_{t-1} - \gamma_t(q_t) \} \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{ \phi_s Q_s - \gamma_s(q_s) \}.$$

If the monopolist chooses not to continue offering BP provisions, then the rebates paid to buyers from periods prior to t are determined by equation (5). This equation, together with the fact that

$$\sum_{s=k}^{t-1} \beta_s^k = \rho_{t-1}$$

for all $k \leq t-1$, implies that

$$(A10) \quad \beta_t^t + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \beta_s^t \geq \rho_{t-1} \\ + \beta_t^k + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \beta_s^k$$

for all $k \leq t-1$. After taking expected values of both sides of equation (A8), equations (1) and (A10) can be substituted in to yield

$$E_t G_t \leq \{ \phi_t Q_t - \rho_{t-1} Q_{t-1} - \gamma_t(q_t) \} \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{ \phi_s Q_s - \gamma_s(q_s) \}.$$

Comparing this to (A9) demonstrates that the monopolist's expected discounted cash flow is at least as high using BP provisions as it would be with any other pricing convention.

PROOF OF PROPOSITION 1:

Let Ω_t^* refer to Ω_t when the monopolist follows the plan $\{q_t^*, t = 0, 1, \dots\}$, and let $\tilde{\Omega}_t$ represent these expected cash flows when the monopolist switches to the plan $\{\tilde{q}_s, s = t, t+1, \dots\}$ at time t . Let Q_t^* and \tilde{Q}_t refer to Q_t when the monopolist sets $q_t = q_t^*$ and $q_t = \tilde{q}_t$, respectively, through time t .

Suppose $E_t \tilde{\Omega}_t > E_t \Omega_t^*$. By equation (A9), this implies that

$$(A11) \quad \phi_t \tilde{Q}_t - \gamma_t(\tilde{q}_t) \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{ \phi_s \tilde{Q}_s - \gamma_s(\tilde{q}_s) \} \\ > \phi_t Q_t^* - \gamma_t(q_t^*) \\ + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{ \phi_s Q_s^* - \gamma_s(q_s^*) \}.$$

Equation (A1) can be written as follows:

$$(A12) \quad E_0 \Pi_0 = \{ \phi_0 q_0 - \gamma_0(q_0) \} \\ + E_0 \sum_{s=1}^{t-1} \delta^s \{ \phi_s Q_s - \gamma_s(q_s) \} \\ + E_0 \delta^t \left\{ [\phi_t Q_t - \gamma_t(q_t)] \right. \\ \left. + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} [\phi_s Q_s - \gamma_s(q_s)] \right\}.$$

By (A11) the last right-hand-side expression in (A12) is greatest when the monopolist follows the plan $\{q_s^*, s = 0, 1, \dots, t-1\}$ through the time $t-1$ but switches to $\{\tilde{q}_s, s = t, t+1, \dots\}$ for the remaining time. The first two right-hand-side expressions are not affected by this switch, so it follows that $E_0 \Pi_0$ is also greatest when the monopolist switches to $\{\tilde{q}_s, s = t, t+1, \dots\}$ at time t ; but

this is a contradiction, since $\{q_t^*, t = 0, 1, \dots\}$ maximizes $E_0 \Pi_0$. Hence, $E_t \Omega_t \leq E_t \Omega_t^*$.

Now consider $E_t \Gamma_t$. Note that $E_t \Gamma_t = E_t F_t$, where $E_t F_t$ is defined by equation (A7). The monopolist at time t chooses $\{\tilde{q}_s, s = t, t+1, \dots\}$ to maximize $E_t F_t$. Equation (A7) can be rewritten as

$$\begin{aligned} (A13) \quad E_t \Gamma_t &= E_t F_t \\ &= \{\phi_t(q_t + q_{t-1}) \\ &\quad - \alpha_{t-1} q_{t-1} - \gamma_t(q_t)\} \\ &\quad + E_t \sum_{s=t+1}^{\infty} \delta^{(s-t)} \{\phi_s(Q_s - Q_{t-2}) - \gamma_s(q_s)\}. \end{aligned}$$

Comparison of equations (A13) and (A12) reveals that the plan $\{\tilde{q}_s, s = t, t+1, \dots\}$ that maximizes $E_t \Gamma_t$ is not, in general, the plan that maximizes $E_0 \Pi_0$.

REFERENCES

- Ausubel, Lawrence M. and Deneckere, Raymond J., "One is Almost Enough for Monopoly," *Rand Journal of Economics*, Summer 1987, 18, 255-74.
- and —, "Reputation in Bargaining and Durable Goods Monopoly," *Econometrica*, May 1989, 57, 511-31.
- Belton, Terrence M., "A Model of Duopoly and Meeting or Beating the Competition," Research Papers in Banking and Financial Economics, Board of Governors of the Federal Reserve System, working paper, 1986.
- Bond, Eric W. and Samuelson, Larry, "Durable Good Monopolies with Rational Expectations and Replacement Sales," *Rand Journal of Economics*, Autumn 1984, 15, 336-45.
- Broadman, Harry G. and Montgomery, W. David, *Natural Gas Markets after Deregulation*, Washington, DC: Resources for the Future, 1983.
- Bulow, Jeremy, "Durable-Goods Monopolists," *Journal of Political Economy*, April 1982, 90, 314-32.
- Coase, Ronald, "Durability and Monopoly," *Journal of Law and Economics*, April 1972, 15, 143-49.
- Cooper, Thomas, "Facilitating Practices and Most-Favored-Customer Pricing," unpublished dissertation, Princeton University, 1984.
- , "Most-Favored-Customer Pricing and Tacit Collusion," *Rand Journal of Economics*, Autumn 1986, 17, 377-88.
- Fama, Eugene, "The Effects of a Firm's Investment and Financing Decisions on the Welfare of its Security Holders," *American Economic Review*, June 1978, 68, 272-84.
- Grether, David M. and Plott, Charles, "The Effects of Market Practices in Oligopolistic Markets: An Experimental Examination of the Ethyl Case," *Economic Inquiry*, October 1984, 22, 479-528.
- Grossman, Sanford J. and Hart, Oliver D., "Takeover Bids, the Free Rider Problem, and the Theory of the Corporation," *Bell Journal of Economics*, Spring 1980, 11, 42-64.
- Gul, Faruk, Sonnenschein, Hugo and Wilson, Robert, "Foundations of Dynamic Monopoly and the Coase Conjecture," *Journal of Economic Theory*, June 1986, 39, 155-90.
- Holt, Charles A. and Scheffman, David T., "Facilitating Practices: The Effects of Advance Notice and Best-Price Policies," *Rand Journal of Economics*, Summer 1987, 18, 187-97.
- Hubbard, R. Glenn and Weiner, Robert J., "Regulation and Long-Term Contracting in U.S. Natural Gas Markets," *Journal of Industrial Economics*, September 1985, 35, 71-9.
- Kahn, Charles, "The Durable Goods Monopolist and Consistency with Increasing Costs," *Econometrica*, March 1986, 54, 275-94.
- MacAvoy, Paul, *Price Formation in Natural Gas Fields*, New Haven, CT: Yale University Press, 1962.
- Modigliani, Franco and Miller, Merton, "The Cost of Capital, Corporation Finance and the Theory of Investment," *American Economic Review*, June 1958, 48, 261-97.
- Neuner, Edward J., *The Natural Gas Industry*, Norman: University of Oklahoma Press, 1960.

- Png, I. P. L., "Pricing of Capacity to a Heterogeneous Customer Population," UCLA Graduate School of Management, Working Paper No. 87-4, March 1987.
- Salop, Stephen C., "Practices That (Credibly) Facilitate Oligopoly Coordination," in J. E. Stiglitz and G. F. Mathewson, eds., *New Developments in the Analysis of Market Structure*, Cambridge, MA: MIT Press, 1986.
- Scherer, Frederic M., *Industrial Market Structure and Economic Performance*, 2nd Ed., Chicago: Rand McNally, 1980.
- Setser, Vernon G., *The Commercial Reciprocity Policy of the United States, 1774-1829*, Philadelphia: University of Pennsylvania Press, 1937.
- Stokey, Nancy, "Rational Expectations and Durable Goods Pricing," *Bell Journal of Economics*, Spring 1981, 12, 112-28.

A Schumpeterian Model of the Product Life Cycle

By PAUL S. SEGERSTROM, T. C. A. ANANT, AND ELIAS DINOPOULOS*

This paper presents a dynamic general equilibrium model of North-South trade in which research and development races between firms determine the rate of product innovation in the North. Tariffs designed to protect dying industries in the North from Southern competition reduce the steady-state number of dominant firms in the North, reduce the rate of product innovation, and increase the relative wage of Northern workers. (JEL 411, 111)

In his celebrated "product life cycle" paper Raymond Vernon (1966) argued that many products experience cycles. These products are initially discovered and produced in developed countries (the North), and exported to less developed countries (the South). As the techniques of production become more standardized, production shifts to less developed countries due to lower labor costs. These older products are then exported back to developed countries.

The product-life-cycle hypothesis has attracted considerable attention among international-trade theorists in recent years. In this literature, the rate at which an individual firm discovers and successfully markets new products is either treated as exogenously given (Paul Krugman, 1979; David Dollar, 1986, 1987) or as a "deterministic" function of the firm's expenditures on new product development (Robert Feenstra and Kenneth Judd, 1982; Thomas Pugel, 1982; Barbara Spencer and James Brander, 1983; Leonard Cheng, 1984; Richard Jensen and Marie Thursby, 1986, 1987). Thus, from the

individual firm's perspective, successful product innovation is either effortless or guaranteed by large expenditures on new product development. In contrast, Joseph Schumpeter (1942) stressed that firms compete with each other to successfully introduce new products. The recent industrial-organization literature has followed Schumpeter's lead (see, e.g., Glen Loury, 1979; Tom Lee and Louis Wilde, 1980; Jennifer Reinganum, 1982). In these research and development (R&D) models, there are losers as well as winners because a firm can spend substantial resources on new product development only to find that another firm has discovered and patented the new product first.

In this paper, we construct a dynamic, general equilibrium model of North-South trade that combines the product-life-cycle hypothesis with Schumpeter's (1942) description of product innovation. We model each R&D race as an "invention lottery" in which the probability of winning the race is proportional to resources devoted to R&D by each firm. The duration of each R&D race is a deterministic decreasing function of the amount of aggregate resources devoted to R&D. Every time a new product is discovered, a new R&D race between firms in the North begins. The winner of each R&D race earns dominant firm profits for an exogenously given patent period, after which perfect competition prevails. Firms in the North choose how much labor to hire for R&D by maximizing expected discounted profits, and consumers maximize their discounted lifetime utility.

*Department of Economics, Michigan State University, East Lansing, MI 48824; Department of Economics, Delhi School of Economics, Delhi University, New Delhi, India; and Department of Economics, University of Florida, Gainesville, FL 32611, respectively. We thank Richard Brecher, Robert Feenstra, Gene Grossman, Paul Krugman, Dave Richardson, Barbara Spencer, Marie Thursby, and the other participants at the NBER summer institute on international studies, July 1987, for helpful comments. The paper has also benefited from the comments of Earl Grinols, Haideh Salehi-Esfahani, and three anonymous referees. Any errors that remain are our own responsibility.

We show that a unique steady-state equilibrium exists in which the number of new products, consumer expenditures, and assets are all constant over time. In the steady state, Northern workers earn higher wages than their Southern counterparts if the South has a sufficiently large fraction of the world labor force. Moreover, the pattern of trade continuously changes with each product initially being exported and then later imported by the North.

Endogenizing the rate of technological change generates some surprising comparative steady-state results in our model. When wages in the North and in the South are equal, an increase in the patent length (or a decrease in the rate of technology transfer to the South) increases the rate of product innovation in the North. This result is consistent with the partial-equilibrium industrial-organization literature on R&D competition, because an increase in the patent length increases the reward for winning an R&D race. However, when Northern workers earn higher wages than Southern workers, an increase in the patent length *decreases* the rate of product innovation in the North. The increase in the patent length, by itself, increases the reward for innovative activity, but Northern wages rise more than enough to offset this effect.

Unlike the previously mentioned studies, we also examine the effects of tariffs designed to protect dying industries in the North from Southern competition. When Northern workers earn higher wages than Southern workers, we find that an increase in the number of industries being protected in the North leads to higher relative wages for Northern workers and a slower rate of innovation. Thus, we are able to theoretically link protectionist trade policies with slower economic growth.

The rest of this paper is organized as follows: In Section I, the dynamic general equilibrium model of North-South trade is presented. In Section II, we characterize the steady-state equilibrium of the model. The relationship between relative labor endowments and steady-state relative wages is explained in Section III. Section IV analyzes the effects of changes in patent length

and tariff protection on the steady-state equilibrium. Finally, our conclusions are presented in Section V.

I. The Model

Consider a world consisting of two countries: the North and the South. Let \bar{L}^N and \bar{L}^S be the aggregate endowments of labor in the North and in the South, respectively. These labor endowments do not change over time.

In this world, there is a countably infinite set of product $N \equiv \{1, 2, 3, \dots\}$. At any point in time $t \in [0, \infty)$, these products can be partitioned into three sets: the set of products that any firm in the world knows how to produce, the set of products that only one firm in the world knows how to produce, and the set of products that no firm in the world knows how to produce. Firms that produce products in the second set are called dominant firms.

At time t , every firm that knows how to produce product $j \in N$ has the same production technology. Constant returns to scale prevail in production, with one unit of labor producing one unit of product j . Labor is the only factor of production, and all workers in the world are equally productive. Factors of production are not mobile internationally.

At time $t = 0$, there are n products for which the production technology is common knowledge, and the remainder of the products have unknown production technology. Time $t = 0$ represents the beginning of a sequence of R&D races between firms in the North. Only workers in the North are capable of doing R&D-type work, and therefore only firms in the North compete in these R&D races. Every time an R&D race in the North ends, a new R&D race immediately begins. At the beginning of the j th R&D race, each firm i in the North must decide how much labor L_{ij}^R to devote to R&D. This choice by firm i represents a commitment to employing L_{ij}^R units of R&D labor for the duration of the j th R&D race. The firm that wins the j th R&D race becomes the sole producer in the world of product $n + j$ for a time period of length

$T > 0$. After this "patent" expires, the production technology for product $n + j$ becomes common knowledge.¹

We model each R&D race as an "invention lottery." Each of the L_j^R worker-researchers participating in the j th R&D race is equally likely to discover product $n + j$ at time $\hat{t}_j = h(L_j^R)$ after the beginning of the j th R&D race; where $L_j^R \equiv \sum_i L_{ij}^R$ is the aggregate labor devoted to R&D. Thus, at the end of the j th race, nature draws one of the L_j^R "lottery tickets" and one of the L_j^R worker-researchers discovers the new product. The firm that employs the winner earns dominant firm profits until its patent expires. Firm i wins the j th race with probability L_{ij}^R / L_j^R .

By modeling the R&D process as an "invention lottery," we capture two features of R&D races that we feel are important. First, individual firms investing in R&D face an uncertain return; there are winners and losers. In contrast, with other models of technological change and international trade (see, e.g., Jensen and Thursby, 1986; Gene Grossman and Elhanan Helpman, 1989; Feenstra and Judd, 1982), a firm is not guaranteed success in developing a new product by spending some fixed sum of money on product development. Secondly, new products tend to be discovered faster and at a greater discounted cost, as more resources are devoted to R&D [this is implied by properties of the $h(\cdot)$ function to be specified shortly]. There is an intertemporal trade-off associated with R&D activity. From the individual firm's perspective, as it spends more money on R&D (a flow expenditure), its possibility of winning the race increases, other firm's probabilities of success decline, and the race ends sooner.²

¹ T is inversely related to the rate of technology transfer in Krugman (1979) and Jensen and Thursby (1986). "Patents" need not be given a literal interpretation. The patent length T serves as a proxy for all relevant factors that impede technology transfer between the North and South.

²The deterministic length $\hat{t} = h(L^R)$ of each R&D race is admittedly artificial, but as will become clear, the main results in the paper are driven by factor market constraints, which would be present whether the length of each R&D race were deterministic or

In this model, at each point in time t , wages for workers in each country are determined by competitive market forces. We set the equilibrium wage rate for workers in the South equal to one and let w denote the equilibrium relative wage of workers in the North. Both production workers and R&D workers in the North are paid the same wage w .

In addition to the absence of international labor mobility, we assume that production of goods protected by patents takes place only in the North. In other words, Southern firms can produce a good only after its patent protection has expired. Possible institutional justification for this assumption would be that enforcement of patent laws in the South is considerably weaker than in the North and the labor market within the South constitutes an effective channel of technology diffusion. Thus, in the absence of effective patent protection in the South, a Northern dominant firm producing in the South faces the risk that some of its workers might establish another firm manufacturing the same product.³

Infinitely lived consumers maximize total lifetime utility. Each consumer has an identical time-separable utility function

$$(1) \quad U \equiv \int_0^{\infty} e^{-\rho t} \log u(\cdot) dt$$

where $\rho > 0$ is the constant subjective discount rate and $u(\cdot)$ is an instantaneous utility function.⁴ We adopt a particular

stochastic. To analyze tractably the effects of commercial policy in a general equilibrium setting, we also abstract from certain interesting features of R&D races (established-firm advantages, variable R&D expenditures over time, imitation in spite of patent protection, etc.) that have been extensively studied in the partial-equilibrium industrial-organization literature.

³Allowing dominant firms to produce in the South could generate multinational firms along the lines proposed by Grossman and Helpman (1989). Multinational activity in our model results in the wage being equalized between the North and South unless all Northern labor is engaged in R&D.

⁴The same form of total lifetime utility is used by Grossman and Helpman (1989).

form of $u(\cdot)$,

$$(2) \quad u(x_1, x_2, x_3, \dots) \equiv \prod_{j=1}^n \left(\sum_{i=0}^{\infty} \alpha^i x_{j+ni} \right).$$

This is a generalized symmetric Cobb-Douglas utility function where n can be interpreted as the number of product groups. Since products within each group are perfect substitutes, we call this the CDP (Cobb-Douglas with perfect substitutes) utility function. Product group j ($j=1, 2, 3, \dots, n$) consists of products $j, n+j, \dots$; and $\alpha > 1$ represents the extent to which each new product improves upon existing products in the same product group.

To illustrate the effect of product innovation on consumer utility, suppose that initially there are n products available for consumption. Given time separability, consumers are, in effect, maximizing the utility function $\bar{U} \equiv x_1 x_2 x_3 \dots x_n$ at that instant in time. The discovery of product $n+1$ means that consumers are now, in effect, maximizing the utility function $\bar{U} \equiv (x_1 + \alpha x_{n+1}) x_2 x_3 \dots x_n$. If the equilibrium prices of products 1 and $n+1$ both equal one, which would be the case if both products 1 and $n+1$ were produced competitively, then no consumer would purchase product 1 (given $\alpha > 1$), and it would become obsolete. Thus, new products substitute perfectly for old products, and product innovation in our model takes the form of superior products replacing inferior products.⁵

We assume that there is a capital market in the North which supplies the savings of Northern consumers to firms engaged in R&D. The equilibrium interest rate $r(t)$ clears the capital market at each point in time t . Firms borrow funds from this market to pay workers as the R&D is done. Each firm issues a risky security that yields a positive return if it wins and a negative return if it loses an R&D race. Assuming

risk neutrality and perfect competition in R&D, Northern firms enter each R&D race until expected discounted profits are driven to zero. Southern consumers are not allowed to participate in the capital market, and therefore at each instant of time, their income equals their expenditure. If this assumption is relaxed, then Southern savings would end up financing part of the North's R&D expenditure, a result that is contrary to empirical evidence.⁶

Free trade is assumed to exist between the North and the South throughout time, and products are assumed to be non-storable. Furthermore, at any time t , perfect competition prevails in the market for each product whose patent has expired. Thus, the market price for all such products equals the marginal cost of production in the South (one). Given the consumer preferences, one unit of product j gives each consumer as much utility as α units of product $j-n$. When both products are competitively produced and sell at the same market prices (equal to one), the competitive market for product j renders product $j-n$ obsolete.

The endowment of labor in the North \bar{L}^N is assumed to be sufficiently small so that, even if all the workers in the North did R&D-type work, the number of dominant firms would be less than the number of product groups n . That is, $h(\bar{L}^N)n > T$. This condition guarantees that there are never two dominant firms producing products in the same product group.⁷ At time t , the dominant firm producing product j must

⁶All the comparative steady-state results concerning the effects of changes in labor endowments, patents, and tariffs on the number of dominant firms, the rate of innovation, relative wages, and world assets would be unaffected if the capital market were international. Nor would the magnitudes of these variables be affected. However, the distribution of world assets between the North and the South and the pattern of trade in the steady state would change.

⁷Even if innovations did not occur in the previously described sequence, all the results in the paper would be unaffected if product j is only discovered when product $j-n$ is competitively produced. For example, it is possible that certain groups do not experience any innovation at all.

⁵Nancy Stokey (1988) models product replacement in a different context. The rest of the product-life-cycle literature has treated product innovation as being the introduction of greater variety.

compete only against a competitive fringe of firms producing product $j - n$.

The dominant firm and firms in the competitive fringe simultaneously set prices, and we solve for a Bertrand-type Nash equilibrium. Let E^W denote instantaneous world expenditure. Given the instantaneous CDP utility function, world expenditure at time t on products in product group j is E^W/n . With the equilibrium price of one being charged by firms in the competitive fringe, the dominant firm has zero sales if it charges a price p^d greater than α . On the other hand, the competitive fringe has zero sales if the dominant firm charges a price p^d less than α . If $p^d = \alpha$, then consumers are indifferent between spending E^W/n on product j and spending E^W/n on product $j - n$. We assume that all the indifferent consumers buy from the dominant firm (all rules for rationing the demand of indifferent consumers among firms are somewhat arbitrary). Then dominant-firm profits are

$$(3) \quad \pi^d(p^d) = \begin{cases} 0 & \text{if } p^d > \alpha \\ (p^d - w)E^W/p^d n & \text{if } p^d \leq \alpha. \end{cases}$$

These profits are clearly maximized where $p^d = \alpha$. Thus, in the Nash noncooperative equilibrium in prices that we examine in the rest of this paper, each dominant firm produces $q^d \equiv E^W/\alpha n$ and earns profits

$$(4) \quad \pi^d \equiv \frac{E^W}{n} \left(\frac{\alpha - w}{\alpha} \right).$$

The competitive fringe constrains each dominant firm from charging prices higher than α .

Several restrictions are placed on the $h(\cdot)$ function that defines the R&D technology. First, $h(\cdot)$ is assumed to be continuously differentiable with $h'(\cdot) < 0$. This guarantees that product innovation occurs at a faster rate when firms in the North devote more resources to R&D. Second $\bar{h} \equiv h(0) < +\infty$; that is, some product innovation occurs even if no resources are devoted to R&D. Third, $h(L^R) > 0$ and $h'(L^R) \geq 0$ for

all $L^R > 0$; that is, no matter how much labor is devoted to R&D, innovation never occurs instantaneously. Fourth,

$$(5) \quad \frac{d}{dL^R} L^R (e^{\rho h(L^R)} - 1) > 0.$$

Notice that $\int_{-h(L^R)}^0 w L^R e^{-\rho t} dt = w L^R (e^{\rho h(L^R)} - 1)/\rho$ is the discounted labor cost of developing a new product (discounted to the end of the R&D race) when the market interest rate $r(t)$ equals each consumer's subjective discount rate ρ . We show in the next section that $r(t) = \rho$ in the steady-state equilibrium. Thus, this condition states that the appropriately discounted labor costs of developing a new product rise as firms try to speed up the process by devoting more resources to R&D. Equation (5) will hold if the $h(\cdot)$ function is downward sloping but sufficiently flat. Fifth, we make a technical restriction

$$(6) \quad -h'(0) < \frac{(1 - e^{-\rho T})h(\bar{L}^N)}{\bar{L}^N (e^{\rho h(\bar{L}^N)} - 1)T}$$

which will also hold if the $h(\cdot)$ function is sufficiently flat. As shown in Appendix A, condition (6) is sufficient but hardly necessary for the steady-state equilibrium to be unique.

Finally, we assume that the labor force in the North (\bar{L}^N) is sufficiently large relative to the labor force in the South (\bar{L}^S) so that

$$(7) \quad \frac{n\alpha\bar{L}^N}{\bar{L}^S + \alpha\bar{L}^N} > \frac{T}{\bar{h}}.$$

As will become clear in Section II, inequality (7) guarantees that, in any steady-state equilibrium, aggregate R&D expenditures are strictly positive.

II. The Steady-State Equilibrium

In this section, we show that a unique steady-state equilibrium exists for the dynamic, general equilibrium model of North-South trade. In this steady state, the number of dominant firms m , the relative wage of Northern workers w , the profit flow

of each dominant firm π^d , the aggregate labor devoted to research and development L^R , world expenditure E^W , world wage income I^W , Northern assets A^N , and the market interest rate r are all positive constants over time and are interrelated in several specific ways.⁸

World expenditure E^W , Northern assets A^N , and the equilibrium interest rate r must be consistent with the consumer's savings-consumption decisions over time. Let t_0 represent the beginning of an R&D race where J innovations have already occurred. The representative consumer's discounted future utility from expenditure path $E(t)$, $t \in [t_0, \infty)$ is

$$(8) \quad U = \sum_{i=0}^{\infty} \int_{t_0+if}^{t_0+(i+1)\hat{t}} e^{-\rho t} \log \left[\frac{\alpha^i E(t)^n}{n^n \alpha^m} \right] dt + \Gamma(J, t_0)$$

where $\hat{t} = h(L^R)$ is the length of each steady-state R&D race and m is the number of products produced by dominant firms. Because the m products produced by dominant firms are sold at price $p^d = \alpha$ and the $n - m$ competitively produced goods are sold at price $p^c = 1$, every time an innovation occurs, the consumer's instantaneous utility increases by factor α . From the point of view of future decision making $\Gamma(J, t_0) = \int_{t_0}^{\infty} e^{-\rho t} \log u(\cdot) dt$ is a constant. Appendix B shows that optimal consumer behavior is characterized by a constant expenditure path over time when the steady-state interest rate equals ρ , the consumer's subjective discount rate. Furthermore, it is shown in Appendix B that the relationship among steady-state expenditure E^W , assets A^N and wage income I^W is

$$(9) \quad E^W = \rho A^N + I^W.$$

Each consumer spends his wage income and interest earnings from his assets at each

instant in time. These assets have been accumulated before the economy reaches the steady-state equilibrium. In other words, equation (9) is just a breakdown of steady-state expenditure by income source.

It is perhaps surprising that, although the instantaneous utility function is characterized by periodic jumps (caused by innovations), there nevertheless exists a steady state with constant expenditures and a constant interest rate over time. However, the optimal expenditure path is derived from the *marginal* utility of expenditure function (see Appendix B), which is invariant with respect to jumps caused by innovations.

Wage income in the world consists of income from production work and income from R&D work:

$$(10) \quad I^W = w\bar{L}^N + \bar{L}^S.$$

Notice that equation (10) implies that R&D workers are paid concurrently. Because goods are nonstorable, world GNP must equal world expenditure E^W :

$$(11) \quad E^W = m\pi^d + w(\bar{L}^N - L^R) + \bar{L}^S.$$

The first two terms represent the value of Northern production, and the last term equals the value of Southern production.

Expected discounted profits of firm i in a typical R&D race are

$$(12) \quad -wL_i^R [1 - e^{-\rho h(L^R)}] + \frac{L_i^R}{L^R} \pi^d e^{-\rho h(L^R)} [1 - e^{-\rho T}].$$

Firm i must pay each of L_i^R workers the wage w for the duration $h(L^R)$ of the R&D race. With probability L_i^R/L^R the i th firm wins the race and earns profits π^d until its patent expires. Perfect competition and free entry in each R&D race drives expected discounted profits of each firm to zero. Summing over all firms in the R&D race, we obtain

$$(13) \quad -wL^R [1 - e^{-\rho h(L^R)}] + \pi^d e^{-\rho h(L^R)} [1 - e^{-\rho T}] = 0.$$

⁸World assets equal Northern assets and Southern expenditure equals Southern income, because Southern consumers do not participate in the capital market.

In other words, aggregate profits discounted to the beginning of an R&D race are equal to zero.

Using equations (10), (11), and (13), it can be shown that the value of aggregate assets is

$$(14) \quad A^N = \frac{m\pi^d - wL^R}{\rho} \\ = \frac{\pi^d}{\rho} \sum_{j=1}^m (1 - e^{-\rho jT}).$$

Equation (14) implies that $A^N > 0$. Assets must be positive in the steady state, because Northern consumers saved in the past to finance the innovation process that led to m dominant firms.

We use geometric techniques to derive the solution and perform comparative steady-state analysis. Combining equations (4), (11), and (13) yields

$$(15) \quad m = Z(L^R, w) \\ \equiv \frac{n\alpha}{\alpha - w} - \frac{[w(\bar{L}^N - L^R) + \bar{L}^S](1 - e^{-\rho T})}{wL^R(e^{\rho h(L^R)} - 1)}.$$

With w fixed, Z can be interpreted as the steady-state zero-profit condition expressed in (m, L^R) space. Given the assumptions about $h(\cdot)$ in Section I, the partial derivatives are unambiguously signed: $\partial Z / \partial L^R > 0$, $\partial Z / \partial w > 0$, and $\lim_{L^R \rightarrow 0} Z(L, w) = -\infty$ for any $w \geq 1$.

The function $m = Z(L^R, w)$ increases in L^R for the following reason: from condition (5), when L^R increases, the discounted cost of developing an innovation $L^R(e^{\rho h(L^R)} - 1)$ increases. Equation (13) implies that to maintain zero discounted profits, dominant firm profits π^d , which represent the reward for winning an R&D race, must be higher. For a given wage w , π^d is higher only if world expenditure E^W is higher [eq. (4)], and from equation (11), there must be more dominant firms for world expenditure to be higher. In other words, firms can only afford to devote more resources to R&D if there are more dominant firms earning positive

economic profits and, thus, higher world income and expenditure.

For a given L^R , an increase in w leads to a proportionate increase in π^d by equation (13). From equation (4), world expenditure must increase more than proportionately. Since world wage income $w(\bar{L}^N - L^R) + \bar{L}^S$ increases less than proportionately with an increase in w , equation (11) implies that m must increase. Thus, for a given L^R , $m = Z(L^R, w)$ increases in w . In other words, for firms to justify paying their R&D workers higher wages, world expenditure on their products must increase, and this can only occur if there are more dominant firms earning profits.

To maintain m dominant firms in the steady state, each time a patent expires, a new product must be discovered. Thus, during the period of time T , m new products must be discovered. This generates the steady-state R&D supply function in (m, L^R) space:

$$(16) \quad m = R(L^R) \equiv T/h(L^R).$$

Given the properties of $h(\cdot)$, the R&D supply increases in L^R ; $\partial R / \partial L^R > 0$ and $R(0) = T/h > 0$, because some product innovation occurs even if no resources are devoted to R&D. Finally, $R(\bar{L}^N) = T/h(\bar{L}^N) < n$, which guarantees that the number of dominant firms is less than the number of product groups.

The relative wage w is determined by competitive market forces. Depending on the distribution of labor endowments between the North and South, the steady-state relative wage w can exceed or be equal to one. The relative wage w cannot be less than one, since Northern workers are as productive as Southern workers and only Northern workers can do R&D-type work.

If $w > 1$ in the steady-state equilibrium, then m products are produced exclusively in the North by dominant firms and $n - m$ products are produced exclusively in the South by competitive firms, with one product from each product group being produced at any point in time. By symmetry, for each product that the South produces, the aggregate output is $\bar{L}^S/(n - m)$,

and the equilibrium price is one. Since this production must satisfy world demand, $\bar{L}^S/(n-m) = E^W/n$. Thus,

$$(17) \quad L^R + \frac{m\bar{L}^S}{\alpha(n-m)} = \bar{L}^N$$

whenever $w > 1$. Equation (17) states that all workers in the North are either engaged in R&D or manufacturing for dominant firms. It implicitly defines $m = F(L^R)$. F can be interpreted as the steady-state labor market constraint in (m, L^R) space. Clearly $\partial F/\partial L^R < 0$. Given equation (17), any increase in L^R must be matched by an equal decrease in the aggregate output of dominant firms. This can only happen if both world expenditure and the number of dominant firms decrease. When $w > 1$ in the steady-state equilibrium, the three graphs $Z(\cdot)$, $R(\cdot)$, and $F(\cdot)$ must simultaneously intersect. This case is illustrated by point A in Figure 1.

On the other hand, if a competitive product is produced in the North, it must be that Northern and Southern production workers get paid the same wage $w = 1$; that is, $L^R + mE^W/n\alpha \leq \bar{L}^N$. Then, the aggregate output for the typical Southern-produced product is $\bar{L}^S(n-m)$. However, this production does not have to satisfy demand; that is, $\bar{L}^S/(n-m) \leq E^W/n$. Thus, when $w = 1$, the graphs $Z(\cdot)$ and $R(\cdot)$ must intersect at a point where the labor market constraint $m \leq F(L^R)$ is satisfied. This case is illustrated by point B in Figure 2. It is proved in Appendix A that a unique steady-state equilibrium exists in both cases.

In this steady-state equilibrium, each product $j \in N$ experiences a Vernon-type product life cycle. Once discovered, each product is produced in the North by a dominant firm for a period of length T . Then production shifts to a competitive industry in the South (if $w > 1$). Eventually each product becomes obsolete, and world production ceases.

III. Labor Endowments and Wages in the Steady-State Equilibrium

In this section, we examine the relationship between relative labor endowments and steady-state relative wages. By varying the

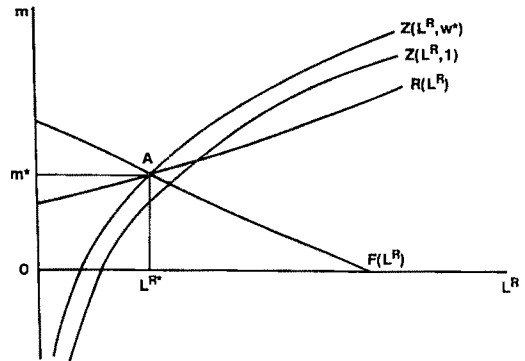


FIGURE 1. THE STEADY-STATE EQUILIBRIUM WITH $w > 1$

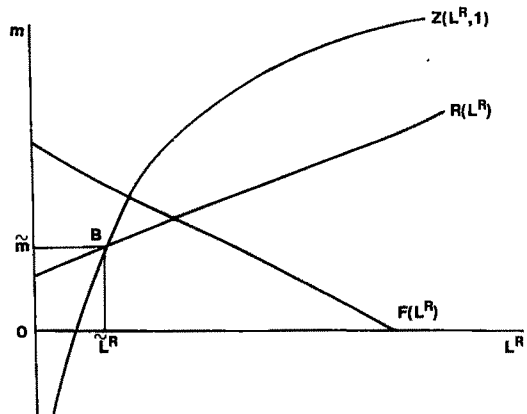


FIGURE 2. THE STEADY-STATE EQUILIBRIUM WITH $w = 1$

Southern labor force \bar{L}^S , we show that if \bar{L}^S is sufficiently small the steady-state relative wage w equals one and that if \bar{L}^S is above some critical value the steady-state relative wage w exceeds one. Furthermore, the comparative steady-state effects of an increase in the Southern labor endowment depend critically on which case we are in. To see this, consider the steady-state effect of a once-and-for-all increase in the Southern labor force \bar{L}^S . Increasing \bar{L}^S causes both the zero profit condition $Z(L^R, 1)$ and the labor market constraint $F(L^R)$ to shift down. If the steady-state relative wage w^* equals one [and eq. (17) holds with strict inequality], then an increase in \bar{L}^S increases the steady-state labor force engaged in R&D (L^R) and increases the steady-state number

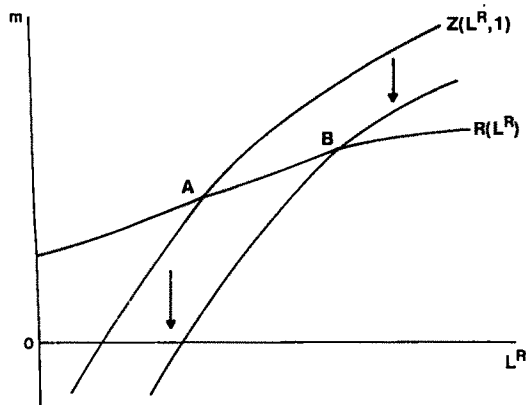


FIGURE 3. EFFECTS OF INCREASE IN SOUTHERN LABOR ENDOWMENT WHEN $w = 1$

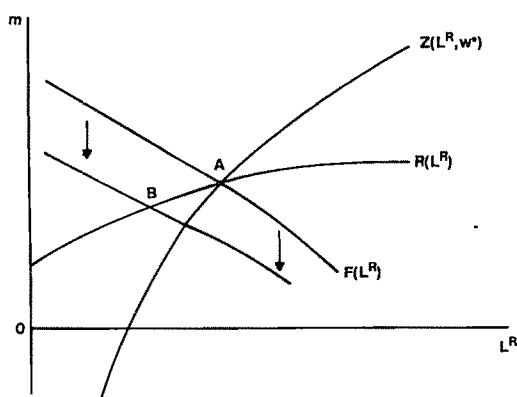


FIGURE 4. EFFECTS OF INCREASE IN SOUTHERN LABOR ENDOWMENT WHEN $w > 1$

of dominant firms in the North (m). This is illustrated by the movement from point A to point B in Figure 3. However, for sufficiently large \bar{L}^S , the labor market constraint becomes binding, and the steady-state relative wage begins to rise. With $w^* > 1$, an increase in \bar{L}^S decreases the steady-state labor force engaged in R&D, decreases the steady-state number of dominant firms in the North, and increases the relative wage of Northern workers.⁹ This case is illustrated by the movement from point A to point B in Figure 4.

The intuition behind this set of results is easy to explain. When the steady-state relative wage w^* equals one and equation (17) holds with strict inequality, at each point in time t , some Northern workers produce competitive products that are also produced in the South. An increase in the Southern labor force \bar{L}^S increases Southern income I^S and Southern expenditure and thus increases world expenditure E^W . As a result of the increase in world expenditure, dominant firms want to produce more ($q^d = E^W/n\alpha$) and to hire more production workers. Since dominant-firm profits [$\pi^d = (\alpha - w)E^W/n\alpha$] increase, perfect competition in each R&D race induces firms to devote more labor L^R to R&D. With a fixed

endowment of labor \bar{L}^N in the North, the increased employment of production workers by dominant firms and the increased employment of R&D labor are exactly balanced by a decreased employment of production workers by competitive firms in the North. However, when the steady-state relative wage w^* exceeds one, this reallocation of labor within the North in response to an increase in \bar{L}^S is not possible, because there are no workers in the North producing competitive goods. Without any change in w^* , an increase in \bar{L}^S leads to excess demand for labor by firms in the North. It is still true that an increase in the Southern labor force \bar{L}^S increases world expenditure E^W , and as a result, dominant firms want to hire more production workers ($q^d = E^W/n\alpha$). Thus, the relative wage w^* must rise enough [and dominant-firm profits $\pi^d = (\alpha - w)E^W/n\alpha$ fall enough] so that firms in the North hire fewer R&D workers; when firms hire fewer R&D workers, this leads to a new steady-state equilibrium with fewer dominant firms.

IV. The Effects of Patents and Tariffs

First, consider the steady-state effect of a once-and-for-all increase in the patent length T (or a decrease in the rate of technology transfer to the South). Increasing T causes the zero profit condition $Z(L^R, 1)$ to shift down and causes the R&D supply function $R(L^R)$ to shift up but leads to no

⁹An increase in \bar{L}^S , given w , shifts down $Z(\cdot)$ (not shown in Fig. 4). Since the final equilibrium is at point B, w and $Z(\cdot)$ must appropriately shift up.

change in the labor market constraint $F(L^R)$. If the steady-state relative wage w^* equals one [and eq. (17) holds with strict inequality], then an increase in T increases both the steady-state labor force engaged in R&D and the steady-state number of dominant firms in the North. This case is illustrated in Figure 5. Increasing T increases the reward for innovative activity. Firms respond to this incentive by increasing the resources they devote to R&D. However, if the steady-state relative wage w^* exceeds one, then an increase in T decreases the steady-state labor force engaged in R&D, increases the steady-state number of dominant firms in the North, and increases the relative wage of Northern workers. This is illustrated in Figure 6. This counterintuitive result is explained as follows: increasing T increases the demand for production workers by dominant firms in the North, because dominant firms have longer lives. Since \bar{L}^N is fixed, the increased employment of production workers in the North must be exactly balanced by a decreased employment of R&D workers. The relative wage w^* must rise enough and dominant-firm profits π^d fall enough so that profit-maximizing firms in the North appropriately reduce their R&D expenditures.

Because each product experiences a Veron-type product life cycle in the steady-state equilibrium, the international trade pattern repeatedly changes over time. Industries in the North die in the sense that production of particular products ceases. Other industries in the North are born when new products are discovered. Thus, in this steady-state equilibrium, production workers in the North repeatedly lose their jobs and must find employment in other sectors of the economy. Given this scenario, tariffs designed to save the jobs of production workers in dying industries would have considerable political support.

We now relax the previous assumption that free trade prevails between the North and the South throughout time and explore the comparative steady-state effects of tariffs designed to protect dying industries in the North from Southern competition. We will assume that the labor force \bar{L}^N in the

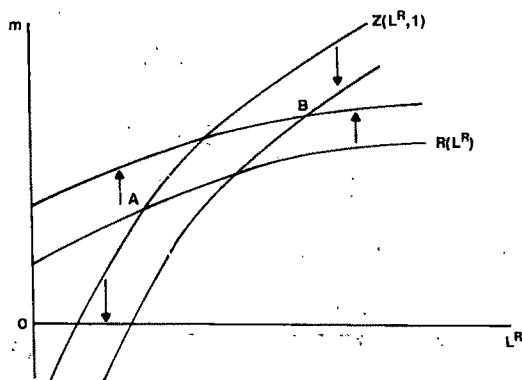


FIGURE 5. EFFECTS OF INCREASE IN PATENT LENGTH WHEN $w^* = 1$

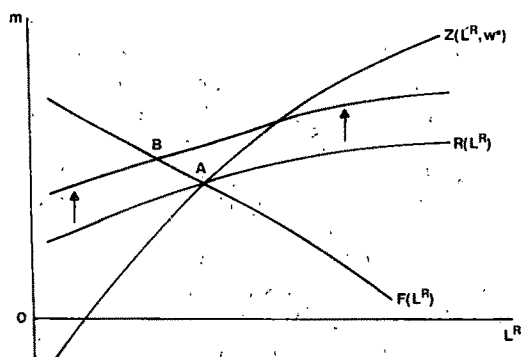


FIGURE 6. EFFECTS OF INCREASE IN PATENT LENGTH WHEN $w^* > 1$

North is sufficiently small so that the steady-state relative wage of Northern production workers w^* exceeds one. This assumption is supported by empirical evidence.¹⁰ At each time t , the government in the North imposes per-unit tariffs on the importation of the \hat{m} products whose patents have most recently expired. The government chooses \hat{m} as its policy instrument; that is, it chooses how many industries in the North to protect from Southern competition. For these tariffs to have any

¹⁰Keith Maskus (1989) and Christopher Clague (1988) present evidence showing that the wage of unskilled workers in some LDC's is about one-tenth that of unskilled Northern workers.

effects on employment in the \hat{m} Northern industries, each per-unit tariff must be at least $w^* - 1$. In other words, each tariff must be prohibitive. We will assume that this holds for each protected industry. As a result, there is no international trade in any of these \hat{m} protected products, and no government revenue is generated by the tariffs.

In the steady-state equilibrium with \hat{m} products protected, m products are produced exclusively in the North by dominant firms, \hat{m} products are produced both in the North and in the South, and $n - m - \hat{m}$ products are produced exclusively in the South by competitive firms. Since Southern income I^S still equals \bar{L}^S in equilibrium, the South must produce \bar{L}^S/n units of each of the \hat{m} protected products for domestic consumption. Since Southern production must satisfy world demand for each of the $n - m - \hat{m}$ products, by symmetry,

$$(18) \quad \frac{E^W}{n} = \frac{\bar{L}^S - \hat{m} \left(\frac{\bar{L}^S}{n} \right)}{n - m - \hat{m}}$$

must be satisfied. The right-hand side of equation (18) represents how much Southern labor must be used to produce each of the $n - m - \hat{m}$ products that are produced exclusively in the South.

Steady-state equations (15) and (16) remain unchanged by the introduction of tariffs. Given that $w > 1$, equation (17) becomes

$$(19) \quad L^R + \frac{mE^W}{n\alpha} + \frac{\hat{m}E^N}{nw} = \bar{L}^N.$$

Using the identity $E^W = E^N + I^S$ and substituting equation (18) into (19) we get

$$(20) \quad L^R + \left[\frac{m}{\alpha} + \frac{\hat{m}}{w} \right] \frac{\bar{L}^S \left(1 - \frac{\hat{m}}{n} \right)}{n - m - \hat{m}} - \frac{\hat{m} \bar{L}^S}{n w} = \bar{L}^N.$$

This equation implicitly defines the new $m \equiv F(\bar{L}^R, \hat{m}, w)$ function. It is easily veri-

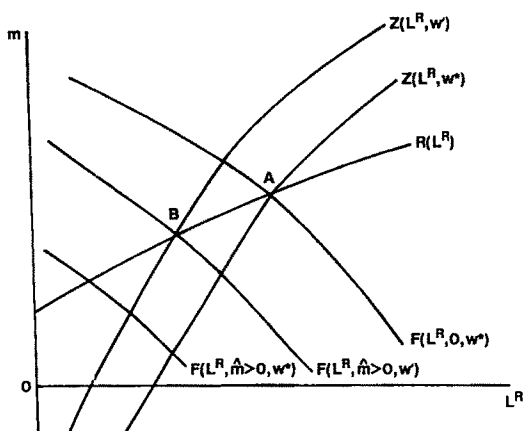


FIGURE 7. EFFECTS OF INCREASE IN THE NUMBER OF PROTECTED INDUSTRIES WHEN $w > 1$

fied that $\partial F / \partial L^R < 0$, $\partial F / \partial \hat{m} < 0$, and $\partial F / \partial w > 0$. Furthermore, when $\hat{m} = 0$, the new $F(\cdot)$ function coincides with the old $F(\cdot)$ function.

Suppose that in the initial steady-state equilibrium $\hat{m} = 0$ and $w = w^* > 1$. This is illustrated by the intersection of the $m = Z(L^R, w)$, $m = R(L^R)$, and $m = F(L^R, \hat{m}, w)$ functions at point A in Figure 7. An increase in \hat{m} shifts down the labor market constraint $F(\cdot)$. Now there is an excess demand for Northern labor at point A. The steady-state relative wage w must rise above w^* , shifting up both the zero profit condition $Z(\cdot)$ and $F(\cdot)$ until a new intersection is established (at point B, with $w = w' > w^*$).

Thus, we can conclude that when Northern production workers earn higher wages than their Southern counterparts, an increase in the number of protected industries in the North (\hat{m}) decreases the steady-state rate of product innovation in the North, decreases the steady-state number of dominant firms in the North, and increases the steady-state relative wage of Northern production workers.

The intuition behind this result is easy to explain. An increase in the number of protected industries in the North raises the demand for Northern production workers. With the Northern labor endowment fixed,

the labor devoted to R&D must decline. This necessitates an increase in Northern relative wages, which raises production costs for dominant firms and lowers dominant-firm profit flows. As a result, R&D becomes less profitable, and the steady-state number of dominant firms (winners of R&D races) declines. Notice that in the case of $w = 1$, increasing the number of protected industries in the North does not have any effect on the steady-state number of dominant firms and the rate of product innovation.

The introduction of a nontraded sector in the North does not affect the impact of protectionism on the rate of product innovation. The effects of patents on the rate of innovation are robust to the introduction of a nontraded good but may not be robust to the introduction of many nontraded goods. If a sufficiently large fraction of the economy involves nontraded goods, then an increase in T could lead to an increase in L^R when $w > 1$.¹¹

V. Conclusion

We have analyzed a dynamic model of innovation, technology transfer, and international trade. Although highly stylized and in some respects unrealistic, this model nevertheless captures some of the forces that are shaping the pattern of trade in the real world today—forces that are not easily captured in the traditional Heckscher-Ohlin trade model.

In our model, sustained product innovation in the North enables Northern workers to earn higher wages than comparable workers in the South. We have carefully analyzed how the rate of product innovation and the relative wage of Northern workers are affected not only by changes in the rate of technology transfer and the world labor endowment, but also by protectionist government policies. What results is an explanation for the small number of new in-

dustries in the North that have arisen to replace older, dying industries as employers of Northern labor. By artificially inflating the wages of Northern workers, protectionist government policies induce sluggish innovative performance in the North.

By focusing on the steady-state equilibrium, we have necessarily abstracted from examining the welfare implications of comparative steady-state exercises. This important task constitutes a nontrivial extension of our model, and it represents a topic for future research.

APPENDIX A

Existence and Uniqueness of Steady-State Equilibrium

Proving that a steady-state equilibrium exists reduces to showing that either (i) $m = Z(L^R, w)$, $m = R(L^R)$, and $m = F(L^R)$ simultaneously intersect at some point $(\tilde{L}^R, \tilde{m}, \hat{w}) \in R_+^3$ for some $\hat{w} > 1$ or (ii) $m = Z(L^R, 1)$ and $m = R(L^R)$ intersect at some point $(\tilde{L}^R, \tilde{m}) \in R_+^2$ where $\tilde{m} \leq F(\tilde{L}^R)$. The graph of $R(\cdot)$ is upward sloping in L^R , and the graph of $F(\cdot)$ is downward sloping in L^R . Moreover, $R(0) < F(0)$ and $R(\bar{L}^N) > F(\bar{L}^N) = 0$ [given the properties of the $h(\cdot)$ function]. Consequently the functions $R(\cdot)$ and $F(\cdot)$ must have a unique intersection, which we will denote (L^{R*}, m^*) . (See Fig. 1).

Suppose that the intersection of $R(\cdot)$ and $F(\cdot)$ lies above the $Z(\cdot)$ graph evaluated at $w = 1$ [$Z(L^{R*}, 1) < m^*$]. Then, increasing the wage shifts the $Z(\cdot)$ graph upward, and since $\lim_{w \rightarrow \infty} Z(L^{R*}, w) = +\infty$, there exists a wage $w^* > 1$ such that all three graphs intersect simultaneously at (L^{R*}, m^*, w^*) . This intersection is unique, corresponds to the case-(i) steady-state equilibrium, and is illustrated by point A in Figure 1.

Suppose that the intersection of $R(\cdot)$ and $F(\cdot)$ does not lie above the $Z(\cdot)$ graph evaluated at $w = 1$ [$Z(L^{R*}, 1) \geq R(L^{R*}) = m^*$]. Notice that $R(0) - Z(0, 1) > 0$ and that $R(L^{R*}) - Z(L^{R*}, 1) \leq 0$ [$\lim_{L^R \rightarrow 0} Z(L^R, 1) = -\infty < 0 < R(0) = T/h$]. This guarantees that the $Z(\cdot)$ function evaluated at $w = 1$ and the $R(\cdot)$ function must intersect for

¹¹An appendix containing an algebraic analysis of the effects of patents and tariffs in the presence of nontraded goods is available from the first author upon request.

some \bar{L}^R and \bar{m} which are less than or equal to L^{R*} and m^* , respectively. Consequently $w = 1$, \bar{L}^R and \bar{m} are steady-state equilibrium values satisfying case (ii). This steady-state equilibrium is illustrated by point B in Figure 2.

The steady-state equilibrium is unique provided that the functions $Z(L^R, 1)$ and $R(L^R)$ have at most one intersection in the interval $(0, \bar{L}^N]$. It suffices to show that

$$(A1) \quad \frac{\partial Z(L, 1)}{\partial L} = [(\bar{L} - L)(1 - e^{-\rho T})Lh'(L)\rho e^{\rho h(L)} + \bar{L}(1 - e^{-\rho T})(e^{\rho h(L)} - 1)] \times [L^2(e^{\rho h(L)} - 1)^2]^{-1} > \frac{\partial R(L)}{\partial L} = \frac{-Th'(L)}{[h(L)]^2}$$

for all $L \in (0, \bar{L}^N)$. From equation (5), it follows that

$$(A2) \quad 0 > e^{\rho h(L)}\rho Lh'(L) > 1 - e^{\rho h(L)}.$$

Substituting (A2) into (A1), it suffices to show that

$$(A3) \quad \frac{1 - e^{\rho T}}{L(e^{\rho h(L)} - 1)} > \frac{-Th'(L)}{[h(L)]^2}$$

for all $L \in (0, \bar{L}^N)$. Since $h''(L) \geq 0$ and $h'(L) < 0$ for all $L \in (0, \bar{L}^N)$,

$$(A4) \quad \frac{-Th'(L)}{[h(L)]^2} \leq \frac{-Th'(0)}{[h(\bar{L}^N)]^2}.$$

Also, by equation (5)

$$(A5) \quad \frac{1 - e^{-\rho T}}{L(e^{\rho h(L)} - 1)} \geq \frac{1 - e^{-\rho T}}{\bar{L}^N(e^{\rho h(\bar{L}^N)} - 1)}$$

for all $L \in (0, \bar{L}^N)$. Combining equations (A3), (A4), and (A5) yields equation (6). Thus, the steady-state equilibrium is unique.

APPENDIX B

Steady-State Consumer Behavior

With time-separable utility, the representative Northern consumer's maximization problem can be solved in two stages.¹² First, for given total expenditure at time t , $E(t)$, and prices of available products, we find the allocation of expenditure that maximizes the consumer's CDP utility function. Then we solve for the time path of expenditures that maximizes U .

The first stage of the consumer optimization problem was analyzed in detail in Section II. The second stage involves choosing the optimum expenditure path $E(t)$, $t \in (0, \infty)$. The representative Northern consumer's assets $A(t)$ evolve according to the equation¹³

$$(B1) \quad \dot{A}(t) = r(t)A(t) + I(t) - E(t)$$

where $r(t)$ is the instantaneous interest rate and $I(t)$ is the consumer's income at time t . Dots denote time derivatives. Furthermore, assets and income must satisfy the feasibility condition

$$(B2) \quad \lim_{t \rightarrow \infty} \inf \left[A(t) + \int_t^\infty \exp \left(- \int_t^\tau r(x) dx \right) \times I(\tau) d\tau \right] \geq 0.$$

This feasibility condition states that the sum of assets and the discounted value of income is nonnegative in the limit as t approaches infinity.

Infinitely lived consumers maximize total lifetime utility [eq. (1)] subject to equations (B1) and (B2). After some algebraic manipulation, equation (8), which is the lifetime

¹²See Hal Varian (1984 p. 148).

¹³See Kenneth Arrow and Mordecai Kurz (1970 Ch. 7).

utility function, reduces to

$$\begin{aligned}
 (B3) \quad U = & n \int_{t_0}^{\infty} e^{-\rho t} \log E(t) dt \\
 & + \left(\frac{e^{-\rho t_0}}{\rho} \right) \log \frac{1}{n^n \alpha^m} \\
 & + (\log \alpha) \left(\frac{e^{-\rho t_0}}{\rho} \right) \left(\frac{e^{-\rho t}}{(1 - e^{-\rho t})} \right) \\
 & + \Gamma(J, t_0).
 \end{aligned}$$

Notice that the second, third, and fourth terms are all constants from the point of view of the consumer [who is choosing $E(t)$]. The third term represents the discounted value of all future innovations in the steady state. If $\alpha = 1$, then new products are identical to old products, and this term disappears.

We conjecture that, in the steady state, the market interest rate $r(t)$ is constant over time and equal to the consumer's discount parameter ρ . We will subsequently verify that, given $r = \rho$, the optimum path of consumer expenditures and assets will also be constant over time in the steady state. Furthermore all markets will clear at each instant in time.¹⁴

Consequently, the consumer solves the following optimal control problem in the steady state at time t_0 :

$$\begin{aligned}
 (B4) \quad \max_{E(t) \geq 0; t \in [t_0, \infty)} & n \int_{t_0}^{\infty} e^{-\rho t} \log E(t) dt \\
 & + \text{constant}
 \end{aligned}$$

subject to $A(t_0) = A_0$,

$$(B5) \quad \dot{A}(t) = \rho A(t) + I - E(t)$$

and

$$(B6) \quad \liminf_{t \rightarrow \infty} \left[A(t) + \frac{I}{\rho} \right] \geq 0.$$

The constraints (B5) and (B6) follow from (B1) and (B2), assuming that $r(t) = \rho$ for all $t \geq t_0$.

The current-value Hamiltonian for this optimal-control problem is

$$(B7) \quad H = \log E(t) + \lambda(t) \{ \rho A(t) + I - E(t) \}$$

and the necessary conditions are $1/E(t) = \lambda(t)$, $\dot{\lambda}(t)/\lambda(t) = 0$, and equations (B5) and (B6). Thus,

$$(B8) \quad E(t) = E_0.$$

Solving (B5), we get

$$(B9) \quad A(t) = e^{\rho t} \left\{ A_0 + \frac{I - E_0}{\rho} \right\} \frac{E_0 - I}{\rho}.$$

If $A_0 + (I - E_0)/\rho < 0$, then the dominant term in (B9) approaches $-\infty$ as t approaches $+\infty$, and the feasibility condition (B6) is not satisfied. If $A_0 + (I - E_0)/\rho > 0$, then the dominant term approaches $+\infty$, and (B6) is clearly satisfied. However, (B6) would still be satisfied if E_0 were increased slightly. Hence higher consumption at every instant in time would be feasible, and therefore the expenditure path $E(t) = E_0$ would not be optimal. Thus, the optimal expenditure path must satisfy

$$(B10) \quad A_0 + \frac{I - E_0}{\rho} = 0$$

and therefore

$$(B11) \quad A(t) = A_0 \equiv \frac{E_0 - I}{\rho}$$

which obviously satisfies (B6).

To summarize, we have shown that, if the steady-state interest rate $r(t) = \rho$, then the consumer optimizes by choosing a constant expenditure path over time E_0 where

$$(B12) \quad E_0 = \rho A_0 + I.$$

¹⁴Alternatively, one could use (6) and (7), allowing the instantaneous interest rate $r(t)$ to vary over time. In this case, if one assumes that expenditure does not change at the steady state, then $r(t) = \rho$; that is, the interest rate is constant in the steady state. For more details, see Grossman and Helpman (1989), in particular their equation 5.

That is, the consumer spends his wage income and interest earning on his assets, at each instant in time.

REFERENCES

- Arrow, Kenneth J. and Kurz, Mordecai, *Public Investment, The Rate of Return and Optimal Fiscal Policy*, Baltimore: Johns Hopkins Press, 1970.
- Cheng, Leonard, "International Competition in R&D and Technological Leadership: An Examination of the Posner-Hufbauer Hypothesis," *Journal of International Economics*, August 1984, 17, 15-40.
- Clague, Christopher K., "Comparative Costs and Economic Development," unpublished manuscript, University of Maryland, 1988.
- Dollar, David, "Technological Innovation, Capital Mobility and the Product Cycle in North-South Trade," *American Economic Review*, March 1986, 76, 177-90.
- _____, "Import Quotas and the Product Cycle," *Quarterly Journal of Economics*, August 1987, 102, 615-32.
- Feenstra, Robert C. and Judd, Kenneth L., "Tariffs, Technology Transfer and Welfare," *Journal of Political Economy*, December 1982, 90, 1142-65.
- Grossman, Gene M. and Helpman, Elhanan, "Product Development and International Trade," *Journal of Political Economy*, December 1989, 97, 1261-83.
- Jensen, Richard A. and Thursby, Marie C., "A Strategic Approach to the Product Life Cycle," *Journal of International Economics*, November 1986, 21, 269-84.
- _____, and _____, "A Decision Theoretic Model of Innovation, Technology Transfer and Trade," *Review of Economic Studies*, October 1987, 54, 631-47.
- Krugman, Paul, "A Model of Innovation, Technology Transfer and the World Distribution of Income," *Journal of Political Economy*, April 1979, 87, 253-66.
- Lee, Tom and Wilde, Louis L., "Market Structure and Innovation: A Reformulation," *Quarterly Journal of Economics*, March 1980, 94, 429-36.
- Loury, Glenn C., "Market Structure and Innovation," *Quarterly Journal of Economics*, August 1979, 93, 395-410.
- Maskus, Keith E., "A Nonparametric Approach to Testing for International Technological Equivalence in Manufacturing," unpublished manuscript, University of Colorado, 1989.
- Pugel, Thomas A., "Endogenous Technological Change and International Technology Transfer in a Ricardian Trade Model," *Journal of International Economics*, November 1982, 13, 321-35.
- Reinganum, Jennifer F., "A Dynamic Game of R&D: Patent Protection and Competitive Behavior," *Econometrica*, May 1982, 50, 671-88.
- Schumpeter, Joseph A., *Capitalism, Socialism and Democracy*, New York: Harper, 1942.
- Spencer, Barbara J. and Brander, James A., "International R&D Rivalry and Industrial Strategy," *Review of Economic Studies*, October 1983, 50, 707-22.
- Stokey, Nancy L., "Learning by Doing and the Introduction of New Goods," *Journal of Political Economy*, August 1988, 96, 701-17.
- Varian, Hal R., *Microeconomic Analysis*, New York: Norton, 1984.
- Vernon, Raymond, "International Investment and International Trade in the Product Cycle," *Quarterly Journal of Economics*, May 1966, 80, 190-207.

Competition by Choice: The Effect of Consumer Search on Firm Location Decisions

By MARC DUDEY*

This paper relates firm location choice and consumer search. Firms that cluster together attract consumers by facilitating price comparison, but clustering increases the intensity of local competition. I construct a simple model which shows that firms may choose head-on competition by locating together. Under reasonable conditions, this is the only equilibrium outcome. (JEL 026)

If consumers find it more convenient to compare the offerings of firms that are located close together, then firm locations can influence consumer search patterns. As emphasized by Tibor Scitovsky (1950) and in the marketing literature (see, in particular, Paul H. Nystrom [1930] and Richard L. Nelson [1958]), firms should take this into account when choosing location. This paper analyzes firm location choice under the assumption that consumers are limited in their ability to compare prices across locations.

My analysis applies to markets in which consumers visit firms to learn prices; advertising and telephone search are assumed to play a negligible role in transmitting price information. One can picture a shopper who is looking for a pair of shoes to match the color of a particular dress. It could well be easier for her to visit a store than to describe what she wants over the phone, while

the cost to a shoe store of supplying sufficiently detailed advertising about its product line may be prohibitive. One might therefore expect shoe store locations to affect the shopper's search pattern. She could, for example, lower her search costs by obtaining price information from all the shoe stores in one shopping center before visiting the shoe stores at another shopping center. This view of the search process is obviously different from the random search process envisioned by George J. Stigler (1961). As noted by Motty Perry and Avi Wigderson (1986), the random-search approach has considerable appeal if consumers engage in telephone search, but it seems inappropriate if consumers learn prices by visiting firms.

My analysis focuses on how firms choose location when they know how consumer search patterns will be affected by their location decisions. Consumers may be attracted to locations occupied by a relatively large number of firms because they expect a relatively high degree of competition there. As a result, firms may have an incentive to cluster together. On the other hand, to the extent that clustering promotes competition, firms have another opposing incentive to locate apart. In Nystrom's (1930) words,

Stores that sell exactly the same kinds of goods and that are clearly competitive do not necessarily merely divide the business that might have been done if there were but one store. Known competition in itself attracts trade, and people come from farther away...

[pp. 137-8]

*Staff Economist, Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 (present address: Department of Economics, Rice University, Houston, TX 77251). This paper reflects my own views and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or members of its staff. The paper is an extensive revision of Dudey (1986). I thank Larry Benveniste, Sally Davies, and other colleagues at the Federal Reserve Board, Kyle Bagwell, Yan Dudey, Jon Eaton, Hugo Sonnenschein, Herman Quirnbach, and Philip White for encouragement, advice, and discussion. In addition, I am grateful to an anonymous referee for numerous comments and suggestions. My thanks also go to seminar participants at the 1988 Winter Econometric Society Meetings, the University of Iowa, and the Federal Reserve Board. All remaining errors are my own.

It may safely be presumed, however, that there is a limit to the good that can come to the individual store from the clustering of competitive shops. While the group secures a greater total volume than could be secured by widely scattered individual stores, it is quite another question as to whether the individual stores may not suffer distinct losses from competition...

[pp. 146-7]

I develop a location-search game to study the tension between these incentives. The game is played by finitely many consumers and quantity-setting firms (to begin with, some form of limited access to the production or retailing technology is assumed to restrict entry into the industry). For simplicity, firms produce at the same constant marginal cost, and consumers have identical demand functions.

In the game, each firm chooses a location before consumers decide where to shop. It is assumed that locations, like shopping centers, can accommodate more than one firm; in the rest of the paper, I will use the term "shopping center" to describe a location that is occupied by at least one firm. A "shopping plan" for a consumer specifies a shopping center that the consumer will visit (if any) for any distribution of firms across shopping centers. The interpretation is that consumers know where firms are located and have enough time to visit only one shopping center.¹ However, a consumer deciding where to shop cannot directly observe the terms of trade available at any shopping center. This is modeled by assuming that firms choose quantity after consumers have decided where to shop.² At each shopping center, consumers buy at a

price that equates local market demand and supply.

Subgame perfect equilibrium (see Reinhard Selten, 1975) is used to solve the game. This requires that firms play A. Augustin Cournot's (1838) quantity-setting game in postlocation competition.³ It also requires that each consumer's shopping plan maximize the consumer's utility for any distribution of firms across shopping centers, given the manner in which firms choose quantities and given the shopping plans of other consumers. Finally, it requires that each firm's choice of location maximizes the firm's profit, given the location choices of other firms, the consumers' shopping plans, and the manner in which firms select quantities.

The central finding is that, under some reasonable conditions, there is a unique, subgame perfect equilibrium outcome in which all firms locate in the same shopping center (although there are parameter values for which firms do not all locate together). Thus, the model can be used to explain why firms selling very similar or even homogeneous products—for example, gas stations, fast-food restaurants, car dealers, and farmers selling fresh produce—often cluster together. This explanation of why firms may locate together differs from the celebrated story told by Harold Hotelling (1929) in that it focuses on consumer search rather than transportation costs as the force driving the clustering phenomenon.⁴ It suggests that clustering will be more pronounced in markets where consumers learn prices by visiting firms and not via advertising or telephone search.

³Quantity-setting Cournot behavior is assumed for simplicity and concreteness. All propositions in the paper could be reformulated using a more general abstract specification of the form of postlocation competition.

⁴C. D'Aspremont et al. (1979) explain why the Hotelling model is an unsatisfactory explanation of clustering if firms choose prices. When firms locate together in the middle of Hotelling's linear market, price competition forces profits to zero. Consequently, each firm has an incentive to locate apart from its rival and generate local monopoly power. This is also the incentive firms have to locate apart in my model. It is the incentive firms have to locate together that makes the two approaches different.

¹Throughout most of the paper, I assume that consumers are endowed with complete information about where firms are located. This assumption has several possible interpretations. Consumers could learn firm locations in the course of their daily travels or from past shopping experiences. Alternatively, firms might advertise their locations. A model in which firms and consumers choose location simultaneously is presented in Section V, Part B.

²An alternative assumption—that firms choose quantity and consumers make shopping plans simultaneously—is discussed in Section V, Part A.

My findings are related to recent work by Konrad Stahl (1981, 1982, 1987) and Asher Wolinsky (1983). Stahl and Wolinsky study models of firm location choice in which firms sell differentiated products and consumers search for a favorite brand. In these papers, it is argued that firms may have an incentive to cluster if consumers are attracted to locations where a large variety of products is available. However, both authors assume that consumers do not expect lower prices at locations occupied by larger numbers of firms. Consequently, it is not the competition between clustered firms that attracts consumers in their models.

The paper is organized as follows. Section I contains a more formal description of the model outlined above. Section II presents the conditions that ensure a unique subgame-perfect-equilibrium outcome of the model in which all firms locate in the same shopping center. It also presents examples which show that, for some parameter values, firms will distribute themselves across more than one shopping center. Section III demonstrates that versions of the existence and uniqueness results hold if entry costs (instead of limited access to the production or retailing technology) determine the number of firms in the industry.

Section IV shows that versions of the existence and uniqueness results can be obtained if firms locate sequentially instead of simultaneously. Two alternative sequencing assumptions are discussed in Section V. The first is that firms choose quantities and consumers make shopping plans simultaneously, and the second is that firms choose location and consumers decide where to shop simultaneously. Section VI contains some concluding remarks.

I. The Basic Model

The analysis is based on a four-stage game that is played by a collection of firms and consumers. The notation used for the basic model is presented in Table 1. Limited access to the production or retailing technology is assumed to fix the number of firms at n . In the first stage of the game, firms simultaneously choose location. The firms may

TABLE 1—NOTATION FOR THE BASIC MODEL

Variable	Definition
m	number of consumers
n	number of firms
$f(p)$	individual consumer demand at price p
c	constant marginal cost of production
$q^C(x, y)$	Cournot equilibrium quantity at a location with x consumers and y firms
$\pi(x, y)$	Cournot equilibrium profit at a location with x consumers and y firms

share locations, and there are at least n locations, so any configuration of firms within the set of locations is permitted. In the second stage, m consumers with the same demand function learn where firms are located and decide if and where to shop. The demand function f , which maps nonnegative prices into nonnegative quantities, is assumed to satisfy two conditions: (i) it is monotonically decreasing as long as it takes positive values and (ii) the area underneath it is bounded. In the third stage, firms choose quantities. These quantities are obtained by the firms at a positive constant marginal cost of c . In the fourth stage, each shopper learns the terms of trade that are available at the shopping center she decided to visit and makes her purchases.

Terms of trade and consumer payoffs are determined as follows. If, at a shopping center with x consumers, the firms collectively choose a total quantity that does not exceed $xf(0)$, consumers make their purchases at a price that clears the market. If the firms collectively choose a total quantity that exceeds $xf(0)$, price equals zero. A consumer who visits a shopping center receives a payoff equal to the surplus she receives from buying at the price at which trade occurs less her round-trip transportation costs (each location can serve as the residence of one or more consumers). Consumers who do not go shopping receive a payoff of zero.

The solution concept that will be used here (subgame perfect equilibrium) requires that the profit-maximizing firms behave as Cournot competitors at any location attracting at least one consumer. I will assume that

a unique, symmetric Cournot equilibrium exists for any combination of firms and consumers.⁵ Let $q^C(x, y)$ and $\pi(x, y)$ denote the Cournot equilibrium quantity and profit per firm at a shopping center with $x \geq 0$ consumers and $y \geq 1$ firms. These functions are assumed to satisfy three standard requirements. Namely, $\pi(x, y)$ is positive for all positive x and y , $\pi(x, y)$ is monotonically decreasing in y for any positive x , and $yq^C(x, y)$, the aggregate quantity supplied in Cournot equilibrium, is monotonically increasing in y for any positive x .⁶

II. Main Results

Since transportation costs complicate matters and are not needed to drive the clustering phenomenon, it will be both instructive and convenient to examine a simplified version of the model in which consumers incur no transportation costs. The case of positive transportation costs is studied in a related working paper (Dudey, 1989a). In that paper, a metric is defined on the set of locations to measure distances, and consumers incur a constant cost k of traveling each unit of distance. Results that are very similar to Propositions 1–5 below are shown to hold in the case of positive transportation costs if the product of k and the maximum distance between consumers is not too large.

The first aim of this section is to prove the central existence result.

PROPOSITION 1: *There is an equilibrium in which all firms locate in the same shopping center.*⁷

⁵For the case of one firm, the unique, symmetric Cournot equilibrium is the unique solution to the quantity-setting monopolist's profit-maximization problem. Sufficient conditions for the existence of a unique, symmetric Cournot equilibrium for any number of firms can be found in F. Szidarovsky and S. Yakowitz (1977), for example.

⁶These assumptions can be derived using the more primitive (but stronger) hypothesis that f is twice differentiable, $f'' < 0$, and $f(c) > 0$.

⁷Unless otherwise specified, "equilibrium" will refer to subgame perfect equilibrium.

The proof will make use of the following preliminary.

LEMMA 1: *The Cournot equilibrium price at any shopping center that attracts at least one consumer does not depend on the number of consumers. In addition, the Cournot equilibrium quantity and profit per firm is linearly homogeneous in the number of consumers.*

PROOF OF LEMMA 1:

Suppose y firms producing at constant marginal cost c face $x \geq 1$ consumers. The unique, symmetric Cournot equilibrium quantity $q^C(x, y)$ is characterized by the inequality

$$(1) \quad f^{-1} \left[\frac{yq^C(x, y)}{x} \right] q^C(x, y) - cq^C(x, y) \\ \geq f^{-1} \left[\frac{q}{x} + \frac{(y-1)q^C(x, y)}{x} \right] q - cq$$

for all q in $[0, \infty)$. By (1) and the uniqueness of the symmetric Cournot equilibrium quantity,

$$(2) \quad xq^C(1, y) = q^C(x, y).$$

It follows from (2) that the Cournot equilibrium price,

$$(3) \quad f^{-1} \left[\frac{yq^C(x, y)}{x} \right] = f^{-1} [yq^C(1, y)]$$

is independent of x and

$$(4) \quad \pi(x, y) = x\pi(1, y).$$

The practical implication of (3) is that an individual consumer does not need to take into account the search patterns of other consumers when she is deciding where to shop. She only needs to consider where firms are located. This result is used in the following argument.

PROOF OF PROPOSITION 1:

The proof is by construction. Let each firm's strategy specify the same location s .

If there are at least three firms, let consumer strategies specify one of the shopping centers occupied by the largest number of firms. If there are two firms, each consumer visits a shopping center; in case one of the two firms is at s and the other is not, all consumers go to s .⁸ A single firm locating outside s would therefore get no consumers, given that rival firms are located at s .

Now, the assumptions that demand is monotonically decreasing as long as it takes positive values, $\pi(x, y)$ is positive if x and y are positive, and $yq^C(x, y)$ is monotonically increasing in y for any positive x imply that the equilibrium price at a shopping center attracting at least one consumer will be monotonically decreasing in y , given the number of consumers attracted. This observation together with (3) and the zero-transportation-cost assumption implies that a consumer cannot improve her payoff by deviating from the strategy given above.

The main idea behind Proposition 1 is that, if all firms are located together, it will not pay any single firm to move to another location, because consumers will correctly predict that such a firm would charge the monopoly price. Proposition 1 explains why firms selling similar or identical products may locate together even though the result will be an increase in the intensity of local competition.⁹

Notice that, since the location s in the proof of Proposition 1 is arbitrary, there are as many clustering equilibria as locations.

⁸Obviously, such a statement cannot be made if consumers live apart from each other and incur positive transportation costs. Thus, the presence of at least three firms is required to extend Proposition 1 to the case of positive transportation costs.

⁹Warren Skoning, formerly in charge of real estate for Sears' Midwestern division, noted that As a general rule, we [Sears] wouldn't want to locate with two other department stores with the same merchandise lines and pricing, but, on the other hand, it may be best to have a third store in your center than to float away to form the nucleus of another, competing center. [quoted in Milton Brown et al., 1970 p. 191]

This statement is easily understood in the context of the basic model.

One might therefore imagine that some intrinsic characteristic of s makes it a focal point (see Thomas Schelling, 1960 pp. 54–8) or that some agent, a "market organizer," proposes s before the firms locate.

The next proposition shows that, if competition between $y+1$ firms is not much more intense than competition between y firms for certain values of y , then clustering is the only equilibrium outcome of the basic model. Formally, the condition that ensures uniqueness of the clustering outcome may be written as

$$(5) \quad \pi(1, y) < (n/y)\pi(1, y+1)$$

for any y not equal to n that divides n . [Examples of demand functions that result in condition (5) being satisfied or not satisfied are presented immediately after the proof of the next proposition.]

The value of condition (5) in proving that clustering can be the only equilibrium outcome is not hard to understand. Loosely speaking, condition (5) implies that firms do not lose too much by locating together. If firms do not lose too much by locating together and do not take into account the effect of their own location choices on the location choices of other firms, they might see clustering as a way of increasing profits. This is because a firm that locates with a few of its competitors may be able to draw consumers (who are attracted to areas of high competition) away from other rivals. As is shown in the following proposition, the conclusion can be that all firms locate together.

PROPOSITION 2: *If condition (5) holds, then equilibrium requires that all firms locate together.*

PROOF:

Recall from the proof of Proposition 1 that the price at any shopping center attracting at least one consumer is monotonically decreasing in the number of firms at the shopping center. This and the assumption that transportation is costless imply that each consumer should visit a shopping center with the largest number of firms to maxi-

mize her payoff. Given that consumers behave in this manner, any firm in a shopping center that is not occupied by the largest number of firms would earn positive (instead of zero) profit by moving to a shopping center that is occupied by the largest number of firms. Thus, equilibrium requires that each shopping center be occupied by the same number of firms.

Now pick an equilibrium and let y denote the number of firms in each shopping center. Observe that some firm must have no more than $[my/n]$ consumers at its location. By the definition of equilibrium, it must be unprofitable for this firm to locate with another group of firms and attract all the consumers. Hence

$$(6) \quad \pi([my/n], y) \geq \pi(m, y+1).$$

It follows from (4) and (6) that

$$(7) \quad \pi(1, y) \geq (n/y)\pi(1, y+1).$$

Since y divides n , (5) and (7) imply that y equals n . Thus, all firms must be in the same shopping center.^{10,11}

Given a demand function f , one can check whether condition (5) holds. For instance, this condition is satisfied if there are at least

three firms and the consumers' demand function is linear.

Example 1. If $f^{-1}(q) = a - bq$, then

$$(8) \quad \pi(1, y) = \left(\frac{(a-c)^2}{b} \right) \left(\frac{1}{(y+1)^2} \right).$$

Condition (5) is easily verified using (8), under the assumption that $n \geq 3$. Thus, three or more firms facing consumers with linear demand will choose to compete by locating together.

The next two examples explain why condition (5) is assumed in Proposition 2. The first of these shows that condition (5) does not hold for all demand functions, and the second shows that this condition never holds in the special case of two firms. In both examples, firms are able to raise industry profits by avoiding the clustering outcome.

Example 2. Suppose there are six quantity-setting firms that produce at zero marginal cost and four consumers with inverse demand

$$f^{-1}(q) = \begin{cases} (1-q)^8 & q \in [0, 1] \\ 0 & q \in (1, \infty). \end{cases}$$

In addition, suppose three firms, A, B, and C, locate together in one shopping center and three firms, D, E, and F, locate at another shopping center. Consumers visit a shopping center with the largest number of firms; in case two shopping centers are occupied by the largest number of firms, consumers a, b, d, and e visit the shopping center containing firms A, B, D, and E, respectively.

The proof of Proposition 2 explains why the consumers' behavior is consistent with equilibrium. To see that the firms' behavior is consistent with equilibrium and that condition (5) does not hold, it is enough to check that each firm profits as much from being a triopolist facing two consumers as from being a quadropolist facing four consumers. In fact, the symmetric Cournot equilibrium profit in the case of three firms

¹⁰The proposition clearly depends on the assumption that firms use pure strategies. For example, consider a version of the basic model in which there are three locations, three firms that may choose location randomly, and three consumers that live apart from each other. It is easy to show that there is an equilibrium in which each firm chooses each location with probability $1/3$. However, given the admittedly restrictive simultaneous-movement assumption, there is a good reason to exclude mixed-strategy equilibria. Namely, if its rivals use mixed strategies, a firm will have an incentive to delay its own move, which is inconsistent with the assumption of simultaneous movement. Note also that mixed-strategy equilibria leave subsets of firms with an incentive to share information about where they intend to locate.

¹¹The result that equal numbers of firms must occupy each firm-occupied shopping center may place strong restrictions on the distribution of firms across locations even if condition (5) does not hold; for example, if n is a prime number, equilibrium requires that firms either all locate together or all locate apart from each other.

and two consumers is 0.01423, and the symmetric Cournot equilibrium profit with four firms and four consumers is 0.01301 (see also Example 4, below).

Example 3. If there are two firms, condition (5) reduces to

$$(9) \quad \pi(1,1) < 2\pi(1,2).$$

Using (4) and the definition of $\pi(\cdot, \cdot)$, condition (9) is equivalent to

$$(10) \quad f^{-1}(q^C(1,1))q^C(1,1) - cq^C(1,1) < f^{-1}(q^C(2,2))q^C(2,2) - cq^C(2,2).$$

However, inequality (10) cannot hold, since $q^C(2,2)$ is available to the monopolist.¹² Thus, condition (5) implicitly requires the presence of at least three firms.

To construct an equilibrium in which the two firms locate apart, suppose there are two consumers and choose strategies for the firms and consumers that satisfy the following properties: the firms locate apart, both consumers' strategies specify the location occupied by both firms if the firms locate in the same place, and consumer *a* (*b*) locates with firm *A* (*B*) if the firms locate apart. By (3) and the zero-transportation-cost assumption, the behavior specified for consumers maximizes their payoffs. Furthermore, the behavior specified for firms is consistent with equilibrium, since the equivalent conditions (9) and (10) do not hold.

III. Entry Costs

In the basic model of Section I, access to the production or retailing technology was assumed to be limited to n firms. This section allows unrestricted access to the technology, but it assumes that firms entering the market incur an entry cost of $E > 0$. A resulting complication is that the ability of a given number of firms to cover entry costs

depends on how the firms are distributed across locations.

The basic model is easily modified to incorporate unrestricted access to the technology and entry costs. Suppose there is a countable infinity of potential entrants and locations. Each potential entrant either decides not to enter and thereby avoids the positive entry cost or chooses a location and automatically incurs the entry cost. Consumers then decide where to shop, and entrants compete for consumers at each shopping center as in the basic model. Payoffs to entrants and consumers are computed as in Section I, except that the entry cost E is subtracted from each entrant's payoff.

In a clustering equilibrium of the unlimited-entry model, entrants must cover their entry costs, and potential entrants who stay out of the market must find it unprofitable to enter and locate with the clustered entrants. This can be formally restated as

$$(11) \quad \pi(m, n) \geq E \geq \pi(m, n+1)$$

where n represents the number of entrants. As long as $\pi(m, 1) > E$, this requirement is satisfied for some n , because E is positive and the area underneath the demand curve is bounded.

If (11) holds, a clustering equilibrium will exist. To see this, suppose that n firms locate in the same place. By the proof of Proposition 1, consumer strategies (consistent with equilibrium) can be specified so that an entrant has no incentive to locate away from the other, clustered entrants. A very similar argument can be used to show that a potential entrant that does not enter the market has no incentive to locate away from the clustered entrants. This concludes the proof of the following extension of Proposition 1.¹³

¹²A referee pointed out that this is essentially the same "monopoly profits exceed joint duopoly profits" point that underlies the preemptive patenting literature.

¹³It should be noted that an equilibrium may not exist if the largest number of clustered entrants that can cover entry costs as a single cluster is two and if there are even arbitrarily small transportation costs (see Dudey, 1989a). The difficulty is related to the "two-firm problem" discussed in Example 3 and footnote 8. Thus, to obtain an existence result in the case of positive transportation costs, it becomes necessary to

PROPOSITION 3: *If $\pi(m, 1) > E$, then there is an equilibrium in the unlimited-entry model in which all entrants locate in the same shopping center.*

A version of Proposition 2 can also be obtained in a straightforward manner if competition among $n + 1$ firms is not much more intense than competition among n firms. In particular, suppose

$$(12) \quad \pi(1, n + 1) > (1/2)\pi(1, n)$$

where n satisfies (11).¹⁴ Now consider an arbitrary equilibrium outcome. Because of the cost of entry, each shopping center must attract at least one consumer. Hence, following the proof of Proposition 2, each shopping center must be occupied by the same number of firms. In fact, similar reasoning combined with (11) implies that this number equals n or $n + 1$. First, suppose it equals $n + 1$. Then, by (11), the firms in any given shopping center must attract all the consumers to cover their entry cost. Thus, there can be only one shopping center. If there are n firms in each of two or more shopping centers, some shopping center must be attracting no more than $m/2$ consumers. By (4), (11), and (12), the firms in that shopping center are not covering entry costs. It follows that there is only one shopping center. This proves the following version of Proposition 2.

PROPOSITION 4: *If condition (12) holds, then equilibrium in the unlimited-entry model requires that all entrants locate in the same shopping center.*

IV. Sequential Firm Location Choice¹⁵

Although the assumption of simultaneous choice of firm location is plausible for some markets and useful for expository purposes,

impose a requirement implying that at least three clustered entrants can cover entry costs.

¹⁴ Using (8), it is easy to verify that condition (12) can hold when consumers have linear demand.

¹⁵ Jon Eaton suggested the topic of this section.

entry into an industry often occurs sequentially. This section considers a version of the basic model in which entrants locate sequentially instead of simultaneously.

Suppose there are n firms that move in sequence, where a move consists of a choice of whether to enter and (assuming entry) where to locate. Let the firms be indexed by their order of movement (firm i is the i th mover). As in the last section, a firm that elects not to enter saves a positive entry cost of E ; any nonentrant commits its resources elsewhere and does not return. Suppose the set of locations is finite and define the rest of the game as in Section I.

In the sequential-entry model, a firm may be able to use its own location choice to influence the location decisions of other firms. As the following example shows, this can guarantee that firms will avoid clustering even when transportation costs are not an issue.

Example 4. Assume there are three consumers and three firms. Suppose $f^{-1}(q) = e^{-\sqrt{q}}$ firms produce at zero marginal cost, and entering firms incur an entry cost of $E = 0.01$. Then, one can easily check that

$$\pi(1, 1) > \pi(3, 2) > \pi(3, 3) > E.$$

Suppose each consumer visits a shopping center occupied by the largest number of firms; in case the firms locate apart from each other, each consumer visits a different firm. As noted above, this behavior is consistent with equilibrium. If the first two firms locate apart, the last firm will locate away from the first two, since $\pi(1, 1) > \pi(3, 2)$. Of course, if the first two firms locate together, the third firm will locate with the first two. Therefore, the second firm may, in effect, choose between locating in a shopping center with two other firms and attracting all the consumers and locating by itself and attracting one consumer. Since $\pi(3, 3) < \pi(1, 1)$, the second firm will locate away from the first firm if the first firm enters. It follows that all firms will enter and locate apart from each other.

Thus, a clustering equilibrium will not generally exist in the sequential-entry model,

even when consumers incur no transportation costs. However, according to the next proposition, the conditions

$$(13a) \quad \pi(1,1) - \pi(1,2) < (2E/m)$$

$$(13b) \quad \pi(1,y) - \pi(1,y+1) < (E/m)$$

for $y = 2, \dots, [n/2]$ rule out almost all equilibrium outcomes except the clustering outcome. Like (5) and (12), these conditions require that competition among $y+1$ firms not be much more intense than competition among y firms for certain values of y . An example in which the conditions are satisfied is presented after the proof of the proposition.

The proof fundamentally depends on the result that all consumers will visit a shopping center occupied by the largest number of firms. Except for the special case of two entrants, conditions (13a) and (13b) give the last entrant an incentive to create a shopping center that is occupied by more firms than any other. The entry cost ensures that no firm will enter unless it forecasts that it will be in the (only) shopping center occupied by the largest number of firms.

PROPOSITION 5: *If $\pi(m,1) > E$, then there is at least one entrant in the sequential-entry model. If conditions (13a) and (13b) hold and there are at least three entrants, then all entrants locate in the same shopping center.*

PROOF:

The existence of an equilibrium follows from Harold W. Kuhn's (1953) backward induction algorithm. The rest of the proof characterizes the set of equilibria.

There must be at least one entrant, since the last potential entrant, firm n , will certainly enter if its predecessors do not. Firm n would find entry profitable since $\pi(m,1) > E$. Now consider an arbitrary equilibrium outcome. Suppose there are several shopping centers occupied by possibly different numbers of firms. Since an entrant receives a payoff of $-E$ if no consumers are attracted to its location, each shopping center must attract at least one consumer. Each

consumer will visit one of the shopping centers occupied by the largest number of firms. An equilibrium outcome must therefore place the same number of firms at each shopping center.

Suppose this number is $y^* \geq 2$ and assume that there are two or more shopping centers, so that $n/2 \geq y^*$. Let x_{\min} be the smallest number of consumers visiting any one of the shopping centers not occupied by the last entrant. Since the firms in each shopping center must be able to cover their entry costs, $\pi(x_{\min}, y^*) \geq E$. Using (4), this implies that the shopping center occupied by the last entrant cannot be attracting more than $m - [E/\pi(1, y^*)]$ consumers and, hence, that $\{m - [E/\pi(1, y^*)]\}\pi(1, y^*) - E$ is an upper bound on the profit of the last entrant.

Evidently, firm n rejected its option to attract all consumers to a location with $y^* + 1$ firms. Its profit from this option would have been $\pi(m, y^* + 1) - E$. Assuming that firm n is not the last entrant, it must be that $\pi(m, y^* + 1) \leq E$. Since the last entrant must cover its entry cost,

$$(14) \quad \{m - [E/\pi(1, y^*)]\}\pi(1, y^*) - E \geq \pi(m, y^* + 1) - E.$$

If firm n is the last entrant, (14) holds by the definition of equilibrium. Rearranging (14) and applying (4) yields the inequality

$$\pi(1, y^*) - \pi(1, y^* + 1) \geq E/m$$

which contradicts (13b). A similar argument together with (13a) may be used to demonstrate that, if there are at least three entrants, $y^* \geq 2$. Thus, there can be no more than one shopping center.

An example in which three out of four firms enter and locate together is presented below. It demonstrates that early movers may be at a disadvantage.

Example 5. Assume there are 12 consumers and four potential entrants. If $f^{-1}(q) = 1 - q^3$, firms produce at zero marginal cost, and entrants incur an entry cost of 1.56,

then it can be shown that

$$\pi(6, 2) < \pi(12, 3)$$

$$\pi(3, 1) < \pi(12, 2)$$

and that the hypothesis of Proposition 5 is satisfied. A simple backward induction argument (left to the reader) can be applied to demonstrate that, if firm 1 enters, firms 2, 3, and 4 may gang up to steal firm 1's share of the market. In this outcome, firm 1 stays out of the market, and firms 2, 3, and 4 locate together in the same shopping center.

The last example of this section gives a basic reason for limiting the number of potential entrants in Proposition 5. The example shows that, if potential entry is unlimited, there may be no entry at all!

Example 6. Assume that there are countably many potential entrants and that

$$\pi(m, 2) > E > \pi(m, 3).$$

Assume that consumers visit the shopping center that was the first to become occupied by the largest number of firms. Now suppose that each firm uses the following rule to decide if and where to enter: if more than one firm has entered and no shopping center is occupied by more than one firm, enter and locate with the last entrant; if one firm has entered, enter and locate apart from the other entrant; do not enter if no other firms have entered or if at least one shopping center is occupied by more than one firm. It is easy to check that the consumer and firm behavior just described is consistent with equilibrium and results in all potential entrants staying out of the market.

V. Models with Alternative Sequencing Assumptions

Following the classical theory of spatial competition, the basic model of Section I assumes that consumers choose location after firms have located. It departs from the classical theory in assuming that consumers choose location before prices are deter-

mined. In effect, the basic model postulates that consumers know the distribution of firms across shopping centers when deciding where to shop and that firms know local demand when choosing quantity. To elaborate on the role of these postulates, this section compares the basic model with two related models that involve alternative sequencing assumptions.

A. A Model with Simultaneous Choices of Quantities and Shopping Plans

This subsection discusses a reasonable variant of the basic model in which firms choose quantity and consumers decide where to shop at the same time. In effect, firms are unable to observe local demand before making quantity decisions in the variant model.

For any particular specification of m , n , $f(\cdot)$, and c , any equilibrium outcome of the basic model is also an equilibrium outcome of the variant model. To see this, fix a subgame of the basic model that arises after firms have chosen locations l_1, \dots, l_n . It has already been shown that equilibrium strategies of consumers may specify any shopping center that is occupied by the largest number of firms. Suppose consumer j visits the shopping center at $s(j)$, $j = 1, \dots, m$, and let y_{\max} equal the largest number of firms at any shopping center. It has also been argued that the equilibrium strategies of firms require that each firm behave as a "local Cournot oligopolist" given the number of consumers that are present at the shopping center it occupies.

Now consider the variant model in which consumers decide where to shop at the same time that firms choose quantity. Fix the subgame of this variant model that arises after firms have chosen the locations l_1, \dots, l_n . Clearly, a firm playing an equilibrium strategy in the variant model still behaves as a local Cournot oligopolist given the number of consumers that are present at the shopping center it occupies. I claim that it is also consistent with equilibrium for consumer j to visit the shopping center at $s(j)$, for all j . If, for all j , consumer j is visiting the shopping center at $s(j)$ and firms are

behaving as local Cournot oligopolists given this distribution of consumers across shopping centers, any consumer is obviously worse off moving to a location that is not occupied by any firms. If any consumer moves to another shopping center with y^* firms that attracts x^* other consumers, she will find that the price she pays increases from $f^{-1}(y_{\max} q^C(1, y_{\max}))$ to

$$f^{-1}\left(\frac{y^* q^C(x^*, y^*)}{x^* + 1}\right)$$

since firms do not adjust quantity in response to the move.¹⁶ It follows that any equilibrium outcome of the basic model is also an equilibrium outcome of the variant game. In particular, clustering remains consistent with equilibrium in the variant game.

There are, however, additional equilibrium outcomes in the variant game. For instance, consider the variant game with three firms and one consumer. It is easy to verify that there is an equilibrium outcome with the following features. Firms A and B locate together in shopping center 1, and firm C locates in shopping center 2. Firms A, B, and C choose quantities 0, 0, and $q^C(1, 1)$, respectively, and the consumer visits shopping center 2 (the consumer visits a shopping center with the smallest number of firms, and in case of a tie, the consumer visits the shopping center occupied by C). Thus, assuming that firms choose quantity and that consumers make shopping plans simultaneously may result in a larger set of equilibrium outcomes.¹⁷

The additional equilibria rely on a coordination problem which occurs when consumers do not visit a shopping center occupied by the largest number of firms. If they were given the opportunity to do so, consumers might try to disrupt such equilibria by notifying firms of their intention to visit a shopping center occupied by the largest

number of firms.¹⁸ In the above example, if the consumer could convince firms A and B of her intention to visit shopping center 1, the consumer as well as firms A and B would be better off. Thus, one may argue that the additional equilibria are unstable in the sense that they are not "communication-proof."

B. A Model with Simultaneous Choices of Firm Locations and Shopping Plans

This subsection considers a version of the basic model in which consumers are uninformed about firm locations in the sense that no firm expects consumers to react to changes in its own location. In other words, firms and consumers are assumed to choose location simultaneously.

The clustering outcome remains consistent with equilibrium in this simultaneous-location version of the basic model. Suppose that all firms and consumers locate in shopping center s . Clearly no individual firm or consumer has any incentive to move from s , since there are no other firms or consumers outside s . It follows that there is an equilibrium in which all firms and consumers locate together.

In fact, given m , n , $f(\cdot)$, and c , any other equilibrium outcome of the basic model is also an equilibrium outcome of the simultaneous-location model. An equilibrium outcome of the basic model satisfies the following properties: it places equal numbers of firms at each shopping center; each consumer visits a shopping center; and if there are $k \geq 2$ shopping centers, then

$$(15) \quad \pi(m, n/k + 1) \leq \pi(x_{\min}, n/k)$$

where x_{\min} is the smallest number of consumers visiting any shopping center. Now consider an arbitrary equilibrium outcome of the basic model as an outcome of the simultaneous-location model. No consumer has any reason to move to a different shopping center, because the same number of

¹⁶This statement is another direct consequence of basic model assumptions and Lemma 1.

¹⁷An example similar to this one was brought to my attention by Kyle Bagwell.

¹⁸This implicitly assumes that communication is possible; see Joseph Farrell (1985).

firms occupies each shopping center. Given that the consumers do not move, no firm would benefit from moving unless

$$(16) \pi(x_{\max}, n/k + 1) > \pi(x_{\min}, n/k)$$

where x_{\max} is the largest number of consumers visiting any location in the specified outcome. But (4) and (15) imply that (16) cannot hold.

Thus, given m , n , $f(\cdot)$, and c , any equilibrium outcome of the basic model is also an equilibrium outcome of the simultaneous-location model. However, the following example demonstrates that there may be additional equilibrium outcomes in the simultaneous-location model. Suppose that an equal number of firms and an equal number of consumers locate in each of y shopping centers. No consumer has an incentive to move to another shopping center, since each shopping center is occupied by the same number of firms. Furthermore, firms have nothing to gain from moving from a shopping center with $n/y - 1$ other firms to a location with no consumers or to a more competitive shopping center with the same number of consumers and n/y other firms. Thus, there is a different equilibrium outcome associated with every common divisor of m and n .

This multiplicity of outcomes depends, of course, on the assumption that no consumers can react to the location decisions of firms. The logic behind Proposition 2 can be used to show that the clustering outcome may still be the unique equilibrium outcome if some (but not necessarily all) consumers can react to firm locations when deciding where to shop.

VI. Conclusion

Classical models of perfect and imperfect competition are generally based on the assumption that consumers have complete information about the offerings of different sellers. However, in reality, the transmission of such information is usually not costless, and the degree of competition between firms will therefore depend on what firms do to facilitate price comparison by consumers.

This appears to be what Scitovsky (1950) had in mind when he wrote

I believe that the market's perfection depends on the buyer's expertness... [T]he geographical concentration of the expert's market and the grading and standardization of products in such a market should not be considered data, as Marshall did. They are the result of a deliberate effort on the part of producers; and I believe that such an effort will only be made in the expert's market, in response to the expert buyer's demand for easy comparability. [Scitovsky, 1950 p. 49]

The consumer-search literature stemming from Stigler's seminal work removes the assumption of costless price comparison. However, almost all of this literature places the burden of information collection on consumers, while firms simply quote prices when approached by consumers. [An exception is the interesting paper by Gerard Butters (1977), which considers the interplay between informative price advertising by firms and consumer search.]

My findings emphasize that the conventional approach to consumer search, which takes the isolation of firms as given, is a very partial equilibrium analysis. Of course, if local competition between two firms is sufficiently intense, firms may choose to locate apart;¹⁹ but, if local competition is not too intense, firms may cluster to facilitate search by consumers.

¹⁹For example, if firms choose price instead of quantity and produce at constant marginal cost, competition between two or more firms drives profits to zero. Although there are multiple equilibrium outcomes in the zero-transportation-cost price-setting version of the basic model, one of the equilibria involves all firms locating apart (see Dudey, 1989a, b).

REFERENCES

- Brown, Milton P., Applebaum, William and Salmon, Walter J., *Strategy Problems of Mass Retailers and Wholesalers*, Homewood, IL: Irwin, 1970.

- Butters, Gerard, "Equilibrium Distribution of Sales and Advertising Prices," *Review of Economic Studies*, October 1977, 44, 465-91.
- Cournot, A. Augustin, *Recherches sur les Principes Mathématiques de la Théorie des Richesses*, Paris: Hachette, 1838.
- D'Aspremont, C., Jaskold-Gabszewicz, J. and Thisse, J. F., "On Hotelling's 'Stability in Competition'," *Econometrica*, September 1979, 47, 1145-50.
- Dudey, Marc P., "Competition by Choice," mimeo, Department of Mathematics, University of Southern California, July 1986.
- _____, (1989a) "Competition by Choice: The Effect of Consumer Search on Firm Location Decisions," mimeo, Board of Governors of the Federal Reserve System, June 1989.
- _____, (1989b) "Consumer Search and Firm Location Choice with Price Setting Firms," mimeo, Board of Governors of the Federal Reserve System, June 1989.
- Farrell, Joseph, "Communication and Nash Equilibrium," Economics Working Paper 85-10, GTE Laboratories, August 1985.
- Hotelling, Harold, "Stability in Competition," *Economic Journal*, March 1929, 39, 41-57.
- Kuhn, Harold W., "Extensive Games and the Problem of Information," *Annals of Mathematics Studies* 28, Princeton: Princeton University Press, 1953.
- Nelson, Richard L., *The Selection of Retail Locations*, New York: Dodge, 1958.
- Nystrom, Paul H., *Economics of Retailing*, Vol. 2, New York: Ronald, 1930.
- Perry, Motty and Wigderson, Avi, "Search in a Known Pattern," *Journal of Political Economy*, April 1986, 94, 225-30.
- Prescott, Edward C. and Visscher, Michael, "Sequential Location Among Firms with Foresight," *Bell Journal of Economics*, Autumn 1977, 8, 378-93.
- Schelling, Thomas, *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, 1960.
- Scitovsky, Tibor, "Ignorance as a Source of Oligopoly Power," *American Economic Review*, May 1950, 40, 48-53.
- Selten, Reinhard, "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Form Games," *International Journal of Game Theory*, January 1975, 4, 25-55.
- Stahl, Konrad, "Consumer Search and the Spatial Distribution of Retailing," Working Paper 366, Institute of Urban and Regional Development at the University of California, Berkeley, 1981.
- _____, "Differentiated Products, Consumer Search, and Locational Oligopoly," *Journal of Industrial Economics*, September/December 1982, 31, 97-113.
- _____, "Theories of Urban Business Location," in Edwin S. Mills, ed., *Handbook of Regional and Urban Economics*, Vol. 2, Amsterdam: North-Holland, 1987, 759-820.
- Stigler, George J., "The Economics of Information," *Journal of Political Economy*, June 1961, 69, 213-25.
- Szidarovsky, F. and Yakowitz, S., "A New Proof of the Existence and Uniqueness of the Cournot Equilibrium," *International Economic Review*, October 1977, 18, 787-89.
- Wolinsky, Asher, "Retail Trade Concentration Due to Consumers' Imperfect Information," *Bell Journal of Economics*, Spring 1983, 14, 275-82.

To Innovate or Not To Innovate: Incentives and Innovation in Hierarchies

By JAMES DEARDEN, BARRY W. ICKES, AND LARRY SAMUELSON*

Hierarchical organizations often perform poorly in inducing the adoption of innovations. We examine a principal offering contracts to agents who make unobservable effort and adoption-of-innovation choices (yielding moral hazard), who occupy jobs of differing, unobserved productivities (yielding adverse selection), and who engage in a repeated relationship with the principal (causing a ratchet effect to arise). Increasing the rate of adoption of an innovation in such an organization causes the incentive costs of adoption to increase at an increasing rate. Relatively low rates of adoption may then be a response to the prohibitive incentive costs of higher adoption rates. (JEL 021, 110, 620)

The Soviet Union has been chronically plagued by difficulties with the diffusion of innovation. In 1941, for example, Georgii Malenkov reported to the 18th Congress of the Communist Party that

... highly valuable inventions and product improvements often lie around for years in the scientific research institutes, laboratories and enterprises, and are not introduced into products.
[Joseph Berliner, 1987 p. 72]

More recently, Mikhail Gorbachev reported to the 27th Party Congress that

... many scientific discoveries and important inventions lie around for years, and sometimes decades, without being introduced into practical applications.
[Berliner, 1987 p. 72]

The technical achievements of the Soviet Union, including such inventions as the hydrogen bomb, Sputnik, and antisatellite and space technology, have been exemplary. As the quotations above attest, the problem lies not in the technical process of invention but in the adoption of the resulting innovations. Why does the Soviet system consistently produce potentially valuable innovations but consistently fail to induce the use of these innovations?

This paper examines innovation in hierarchical organizations, focusing particularly on process (rather than product) innovations. We draw our motivation and examples from the Soviet Union because they are the most striking, but the analysis can be applied to general hierarchical systems.¹

Our basic finding is that a key obstacle to the adoption of innovations in hierarchical organizations is not the cost of inventing or developing an innovation but, rather, the

*Dearden: Department of Economics, Lehigh University, Bethlehem, PA 18015; Ickes: Department of Economics, Pennsylvania State University, University Park, PA 16802; Samuelson: Department of Economics, University of Wisconsin, Madison, WI 53706. The authors thank Joseph Berliner, Eric W. Bond, Herb Levine, participants at seminars at Penn State, the University of Pittsburgh, the University of Windsor, the University of Pennsylvania, and an anonymous referee for helpful comments. Financial support from The American University, the National Science Foundation, and Hewlett Foundation is gratefully acknowledged, as is the hospitality of the Department of Economics at the University of Illinois, where the third author was a visiting professor during the early stages of this work.

¹For example, F. M. Scherer (1984 part III) finds that the rate of innovation on the part of a firm increases as firm size increases but does so at a decreasing rate. He suggests that the relatively poor innovation performance of large firms "is in turn probably due to organizational problems ... although the precise mechanism of this phenomenon is not clear" (p. 191). We take the distinguishing feature of a hierarchical organization to be that the incentives to adopt innovations or take other actions must be explicitly constructed by a principal or planner rather than provided by the market. A similar characterization is adopted by Michael Riordan (1987).

cost of constructing incentives to induce agents to adopt the innovation once it is available. We show that increasing the rate of adoption of an innovation causes these incentive costs to increase at an increasing rate. Relatively low rates of adoption, such as seen in the Soviet Union, may then be a response to the prohibitive incentive costs of higher adoption rates. Moreover, achieving increased adoption rates without prohibitive cost may require not just a tinkering with the form of incentive contracts, but a modification of the hierarchical decision-making process.²

The first difficulty in constructing incentives to adopt an innovation is that the principal generally cannot observe whether an innovation has been adopted, being instead able to observe only the output of an agent. This is especially likely to be the case with process innovations. Furthermore, the level of output is generally affected not only by the innovation-adoption decision, but also by such factors as an agent's choice of effort, which usually cannot be observed by the principal, as well as unobserved exogenous factors such as the quality of the inputs, facilities, and organization with which an agent must work. We generally refer to these as simply the productivity of the job or enterprise that an agent fills or manages.

A principal desiring to induce the adoption of an innovation must then solve a principal-agent problem with moral hazard (on effort and innovation adoption) and adverse selection (on job productivity).³ Un-

fortunately, a ratchet effect appears in repeated principal-agent relationships with moral hazard and adverse selection.⁴ If the principal is uncertain concerning the productivity of an enterprise or job, the principal has an incentive to use current performance as a signal of that productivity. High current output is then followed by more stringent future remuneration schemes. The benefits from adopting an innovation thus tend to be "ratcheted" away.

In particular, suppose the economy contains high-productivity and low-productivity jobs which cannot be distinguished. The principal will generally seek to induce innovation adoption and high effort, at least from agents in high-productivity jobs. Doing so will require an "innovation-adoption bonus" for the resulting exceptionally high output (denoted y_1) sufficient to induce the agents to bear the cost of adopting the innovation and high effort. Suppose now that agents in low-productivity jobs are also to be induced to adopt an innovation and to supply high effort, yielding an output y_2 with $y_2 < y_1$. An innovation-adoption bonus to cover these costs must now be attached to y_2 . The temptation then arises for agents in the high-productivity jobs to eschew innovation adoption and produce y_2 . These agents can then claim to be occupants of low-productivity jobs who have adopted the innovation and thus collect the adoption bonus without bearing the cost of adopting the innovation. Even more seriously, the agents in high-productivity jobs may adopt the innovation but then mimic low-productivity jobs by supplying low effort. This provides these agents with an innovation-adoption bonus, savings on effort, and a cushion against the ratchet effect arising from having increased the productivity of the job without revealing this increase.

²A substantial literature considers microeconomic models of the diffusion of innovations. Our analysis departs from this literature in *assuming* that the innovation yields revenue increases that exceed direct adoption costs. The difficulty is that the hierarchical nature of the organizations forces a principal to construct costly incentives to induce agents to adopt innovations.

³See Oliver Hart and Bengt Holmström (1987) for a survey of the principal-agent literature. The principal can often come arbitrarily close to a first-best outcome if arbitrarily negative payoffs could be attached to some outcomes (J. Mirrlees, 1974; Holmström, 1979; Steven Shavell, 1979). If ever such a forcing contract could be written, the Soviet Union appears to be the natural place. Even in the Soviet Union, however, there are limits on the penalties that can be imposed. One notes that in the Stalinist period, such limits may

well have not arisen. Interestingly, innovation diffusion rates in the Soviet Union were highest in the 1930's (David Dyker, 1985 p. 28), though we do not wish to argue that the potentially extra severe sanctions were the cause.

⁴See, for example, Berliner (1957), A. Nove (1977), M. L. Weitzman (1980), Holmström (1982), M. Keren et al. (1983), and X. Freixas et al. (1985).

How is this pooling behavior to be deterred and how are the agents in high- as well as low-productivity jobs to be induced to supply high effort and adopt the innovation? The return to output y_1 must be increased even further to make masquerading as a low-productivity job unprofitable. Inducing agents in low-productivity jobs to adopt an innovation thus carries an extra cost related to preserving the desired innovation-adoption incentives for agents in high-productivity jobs. This result readily generalizes to organizations with more than two job productivities. At each step down the scale of job productivities, the adoption of an innovation can be induced only if one pays the direct and incentive costs of adoption to the agents in question and also pays the *increase* in incentive cost to agents in all higher-productivity jobs. This causes the cost of inducing innovation adoption to increase at an increasing rate as one proceeds from high- to low-productivity jobs. The response to this cost-of-adoption schedule may be to induce innovation adoption only in jobs of relatively high productivity and, hence, to induce relatively little use of the innovation.

This result can be contrasted with the outcome of a decentralized or market economy. In the latter, the benefits of an innovation need only exceed the direct costs of adoption in order to induce the agent to innovate. The market will then induce adoption levels that are efficient and that tend to be higher than those of the hierarchical system.

Section I motivates the analysis by providing some evidence on the key features of the model for the case of the Soviet Union. Section II presents a two-period model. Section III presents an equilibrium existence and characterization result. In Section IV we examine potential equilibria and establish their properties. In the process, the workings of the ratchet effect are exposed. Our conclusions are presented in Section V.

I. Innovation and Diffusion in the Soviet Union

Our analysis rests on three stylized facts: that hierarchical systems perform poorly in inducing the use of innovations; that job

productivities, effort levels, and innovation-adoption decisions are difficult to monitor in a hierarchical system; and that the principal's inability to commit gives rise to a ratchet effect. We can illustrate each of these for the case of the Soviet Union.

A growing body of research reveals that the Soviet system of bonus contracts is ineffective in providing innovation incentives (Phillip Hanson, 1981 p. 64; Berliner, 1976) and that differences in innovation are so important as to be a major cause of the technological gap between East and West (Ronald Amann and Julian Cooper, 1986 p. 12).⁵ There is also evidence that the problem lies not with the technological process of invention but with the failure of innovations to diffuse in the Soviet Union. For example, Table 1 reports the date of the first introduction of various innovations in the Soviet Union and several Western economies. Table 2 reports data on the spread of these technologies.

Table 1 indicates that the Soviet Union's record in developing advanced technologies is quite good. The initial dates of commercial production of the various technologies generally lag only slightly behind those of the four Western economies. Table 2, however, reveals that the subsequent diffusion of these technologies into widespread use has proceeded at a much slower pace in the Soviet Union than in the West. In every case, a significantly higher fraction of 1982 output is produced by the new technology in the Western economies than in the Soviet Union. Given the Soviet tendency to concentrate innovation efforts in leading enterprises such as steel and nuclear power, the data in Tables 1 and 2, if anything, overstate the success of Soviet innovation attempts.⁶

⁵A striking illustration of the bonus-contract system's failure to induce adequate diffusion is provided by the fact that some innovations have so stubbornly resisted diffusion as to spread into general use only after direct intervention on the part of the highest political leadership. Inducing the use of natural gas, for example, required the personal efforts of Nikita Khrushchev (Nove, 1977 p. 187).

⁶See, for example, Amann and Cooper (1982 p. 24). Similar diffusion experiences characterize other planned economies. For example, Steven Popper (1988) studies the diffusion of numerically controlled machine

TABLE 1—ADOPTION OF NEW TECHNOLOGIES: DATES OF FIRST COMMERCIAL PRODUCTION

Technology	USSR	USA	Japan	FRG	UK
Oxygen steel	1956	1954	1957	1955	1960
Continuous cast steel	1955	1954	1960	1954	1958
Synthetic fiber (nylon)	1948	1938	1942	1941	1941
High-pressure polythene	1953	1941	1954	1944	1937
Nuclear power station	1954	1957	1966	1961	1956
Numerically controlled machine tools	1965	1957	1964	1963	1966

Source: Amann and Cooper (1986 p. 12).

TABLE 2—SUBSEQUENT DIFFUSION OF NEW TECHNOLOGIES: PROPORTIONS OF OUTPUT PRODUCED BY NEW TECHNOLOGIES IN 1982

Technology	USSR	USA	Japan	FRG	UK
Oxygen steel (as percentage of total steel)	29.6	62.1	73.4	80.9	66.1
Continuously cast steel (as percentage of total steel)	12.1	27.6	78.7	61.9	38.9
Synthetic fiber (as percentage of total man-made fiber)	51.2	91.2	83.8	83.1	78.6
Polymerized plastics (as percentage of total plastics)	46.4	87.5	80.0	73.0	79.3
Energy generated by nuclear power station (percentage of total)	7.1	12.4	17.6	17.3	16.7
NC machine tools (as percentage of total metal-cutting machine tools)	16.6	34.0	52.8	20.6	27.7

Source: Amann and Cooper (1986 p. 13).

Our analysis presumes that job productivity as well as effort and innovation-adoption levels cannot be observed by the principal (or at least cannot be observed without exorbitant cost). While it is natural to think of effort as unobservable, one might conjecture that it is easy to observe whether an enterprise manager has adopted an innovation. However, the evidence suggests otherwise. In order to fulfill innovation-adoption targets, for example, Soviet managers frequently either adopt artificial or "pseudoinnovations" that represent only superficial changes in the process of production

(Berliner, 1976 p. 375) or claim innovation adoptions that are actually nonexistent:

Where the [innovation] plans are fulfilled ... one would expect that the technological level would be satisfactory, but in fact this is not always the case.... This state of affairs was recognized by [Leonid] Brezhnev at the XXV Party Congress when he pointed out that "there are still products which in the reports appear as 'new' but in fact are new only by the date of production and not by their technical level." [M. J. Berry, 1982 p. 82]

tools in Hungary. He finds much longer diffusion lags in Hungary than in Western economies. Neil Leary and Judith Thornton (1989), in a study of the Soviet steel industry, find not only slower diffusion rates but utilization rates that peak at "much lower ceilings than those in the market economies" (p. 65).

We can also provide evidence that Soviet planners cannot observe effort or, more generally, input levels. In the late 1960's, for example, managers of the Shchekino Chemical Plant were allowed to keep any cost savings that could be achieved by employing labor more efficiently and releasing excess

labor for other uses. The response was an increase in labor productivity of 52 percent in the first year. This experience is revealing both because of the extent of the labor hoarding or inefficient input use that persisted under the conventional monitoring system and because the response to suspected labor hoarding was not increased monitoring but revised incentives. This presumably testifies to the difficulty of monitoring.⁷

If anything, it is not even clear that output can be observed. This is evident, for example, in the existence of the "second economy," where finished goods are often diverted from official channels by claims that they are "spoiled" (Gregory Grossman, 1981 p. 76). It is also indicated by the frequency with which output reports are inflated to make performance appear better than it is. These inflated reports go undetected by the conventional monitoring system but are occasionally exposed by extraordinary audits:

...spot checks of 48 enterprises belonging to the USSR Ministry of Construction Materials Industry revealed significant inflated reports at every other enterprise.... Inflated reports were found at 20 out of 24 plants and associations checked in the USSR Ministry of Petrochemical Industry.

[E. Manevich, 1987, pp. 84-85]

Recent Soviet discussions reveal that this problem of inflated performance reports is pervasive.⁸

⁷A somewhat unusual illustration of the difficulties in observing the productivity of an enterprise is provided by noting that Soviet athletes received bonuses for winning medals in the 1988 Olympics (*Time* magazine, October 3, 1988, Vol. 132, No. 14, p. 58). A basic scheme of descending payments for gold, silver, and bronze medals was established. However, an athlete's bonus for winning a particular medal was then adjusted upward if his finish was higher than expected and adjusted downward if the finish was lower than expected. Expected finish thus plays a role much like job productivity. It is imperfectly observed and is based on observations of previous performance.

⁸Increased Soviet attention has recently been focused on the ubiquity of inflated Soviet performance

Our final presumption is that the planner's inability to commit to future remuneration schemes gives rise to a ratchet effect. In the Shchekino experiment, the planning ministry committed to refrain from revising targets for five years. However, the initial gains to managers from the increased labor productivity were quickly dissipated as planners reneged on this "commitment." The Shchekino plant had its instructions rewritten seven times in ten years. A second enterprise operating on the same system suffered 17 changes in five years (Peter Rutland, 1984 p. 353). These are not isolated examples. A recognition of the costs of the ratchet motivated the reform decrees of 1979 and the Andropov Experiment of 1983, which stipulated that the five-year plan was to take precedence over the annual plan in order to lengthen the period of commitment. In practice, however, the planning ministries persisted in continually revising enterprise performance targets (Ed Hewett, 1988 pp. 252, 264-65). The Sibtiashmash Productive Association, for example, had its norm linking wage funds to performance revised four times in 1984 alone (Hewett, 1988 p. 265).

There is ample evidence that this inability to commit gives rise to a ratchet effect. For example:

The Kornevskii Silicate Brick Plant succeeded in 1954 in shortening the autoclave baking cycle to 9.8 hours, while the industry average was 12.4 hours. In 1955 they set its plan at 9.7 hours. Having run into trouble getting enough raw materials, the enterprise failed to fulfill its plan in the first quarter and fell among the lagging enterprises, even though it was producing more per unit of equipment

reports. Spurred by V. Selyunin and G. Khanin (1987), Soviet economists have recognized that inflated performance reports can yield significantly overstated growth rates and understated inflation rates. Khanin, for example, estimates that Soviet national income increased by a factor of 660 percent between 1928 and 1985 as compared to the official figure of 8,900 percent. The controversy generated by Selyunin and Khanin is discussed in R. E. Ericson (1988) and V. G. Trembl (1988).

than other silicate plants which had fulfilled their plans.

[Berliner, 1957 p. 78]

More generally, Yuri Andropov reported that

The business leader who has ... introduced in the enterprise a new technology ... not infrequently is a loser, while those who avoid that which is new lose nothing. [Hewett, 1987 p. 216]

II. A Model of the Incentives to Innovate

A. Extensive Form

We assume that a risk-neutral principal (or central planner) hires or writes a contract with two or more identical risk-neutral agents (or enterprise managers). For convenience, we assume that the relationship between the principal and agents lasts for two periods and the second period is not discounted.⁹

The jobs to be performed by the agents (or enterprises to be managed) can be one of two possible types, either high or low productivity. Agents observe their job's productivity. The principal cannot observe the productivity of a job and must act on the basis of prior beliefs. We can then let the parameter β denote the productivity of a job and let p_1 be the prior probability of high productivity, so that

$$\beta = \bar{\beta} \text{ (high productivity)} \\ \text{with probability } p_1$$

$$\beta = \underline{\beta} \text{ (low productivity)} \\ \text{with probability } 1 - p_1.$$

We can think of nature originally independently choosing a value of β for each job with this value then characterizing the job for *both* periods.

In period one, and after observing β , the agents choose one of two possible effort levels and choose whether to adopt an innovation. Let a denote the choice of effort and θ the innovation-adoption choice, with \bar{a} and \underline{a} denoting high and low effort, respectively, and with $\bar{\theta}$ and $\underline{\theta}$ denoting the choices to adopt and not to adopt the innovation, respectively. The period-one choice of effort level has implications only for period-one output. However, if the innovation is adopted, it makes the job more productive both in periods one and two. The principal and the agents both observe output, and the principal then makes payments to the agents. The principal cannot observe the agents' choices of (a, θ) . The principal updates the principal's prior expectation concerning the productivity of each agent's job based upon the principal's observations of period-one outputs.

The principal now decides whether to have the agents occupy the same jobs in period two as in period one or to transfer them between jobs. If the agents perform the same period-two jobs as they performed in period one, then the accumulation of job-specific human capital causes period-two outputs to be α times the corresponding period-one levels, where $\alpha > 1$. If the agents are transferred, the job-specific human capital is lost. In period two, each job is characterized by the same basic productivity it carried in period one. Agents recall this or observe the productivity of their new job if they have been transferred. If the innovation was adopted in the job in period one, that innovation adoption continues to boost the job's output. If no adoption occurred, no further opportunity arises. Agents then make effort choices, output is realized, and the principal makes a period-two payment to the agents. Table 3 summarizes the sequence of events.

The principal cannot commit to period-two remuneration schemes. Hence, any period-one announcement must include a period-two payment scheme that will be optimal for the principal once period two has arrived. Equivalently, we can think of the principal as announcing the period-two remuneration scheme at the beginning of pe-

⁹The role of the two-period limitation and possible extensions to longer horizons in models of this type are discussed briefly in Ickes and Samuelson (1987).

TABLE 3—SEQUENCE OF EVENTS

Period one:

- 1) Nature chooses productivities of jobs ($\beta = \underline{\beta}$ or $\bar{\beta}$); agents observe β
- 2) Principal announces remuneration scheme for period one and announces whether job transfers will occur
- 3) Agents choose effort levels ($a = \underline{a}$ or \bar{a}) and make innovation adoption choices ($\theta = \underline{\theta}$ or $\bar{\theta}$)
- 4) Outputs are realized and payments to agents made. Principal updates expectations concerning β
- 5) Job transfers occur ($\alpha = 1$) if contract calls for transfers; otherwise, job transfers do not occur ($\alpha > 1$)

Period two:

- 6) Jobs retain values of β and θ chosen in period one
- 7) Principal announces remuneration scheme for period two
- 8) Agents choose effort levels ($a = \underline{a}$ or \bar{a})
- 9) Outputs are realized and payments to agents made

riod two, as shown in Table 3. Given these assumptions about commitment possibilities, we examine a Bayesian subgame perfect equilibrium.¹⁰

The possibility that the agents may be transferred among jobs may appear novel. It has been shown previously (Ickes and Samuelson, 1987) that in the presence of the ratchet effect, optimal contracts may call for regularly transferring agents between jobs. This practice of job transfers removes the incentive for an agent in a high-productivity job to disguise the job's productivity in order to secure more favorable future remuneration schemes. It does so by ensuring that future schemes, applicable to the new job into which the agent has

been transferred, will not depend upon the productivity of the current job.¹¹

Because agents and the jobs they occupy are *ex ante* identical, we can simplify the analysis by hereafter examining the relationship between the principal and a single agent, referred to as "the agent." Job transfers in the single-agent model are equivalent to presuming that job-specific human capital does not appear and that, in period one, the agent takes the principal's expectations concerning the job as exogenous and unaffected by the agent's actions (because they describe expectations in a new job to which the agent is transferred). We will also occasionally refer to the agent in a high-productivity job and the agent in a low-productivity job, but this refers to the two possible types of the job filled by the single agent.

B. Assumptions

A function $y: \{\underline{\beta}, \bar{\beta}\} \times \{\underline{a}, \bar{a}\} \times \{\underline{\theta}, \bar{\theta}\} \rightarrow \mathbb{R}$ gives output levels. The function y is as

¹⁰We assume that the principal can make a credible commitment to transferring agents between jobs. Because transfers are readily defined and verified, it is relatively easy to write and enforce an explicit contract or sustain an implicit contract to transfer employees from one job to another. In contrast, it is likely to be impossible to write and enforce contracts specifying future remuneration schemes. As noted in Ickes and Samuelson (1987), one readily finds examples in which employers commit to transferring employees between jobs, but one rarely finds explicitly specified criteria for evaluation of job performance and remuneration. For example, compare the number of academic departments that commit to times at which tenure reviews will be conducted with the number that explicitly state criteria for tenure.

¹¹We assume that the principal cannot make the decision whether to transfer an agent to a new job contingent upon the agent's output in the current job and cannot contract to transfer an agent into a job of a particular productivity. It is then convenient (and sacrifices no generality) to assume a random-assignment rule. Allowing transfers to depend upon current output alters some of the details of the calculations and equi-

sumed to satisfy

$$(1) \quad y(\bar{\beta}, \bar{a}, \bar{\theta}) \equiv y_1$$

$$y(\bar{\beta}, \underline{a}, \bar{\theta}) = y(\bar{\beta}, \bar{a}, \underline{\theta}) = y(\underline{\beta}, \bar{a}, \bar{\theta}) \equiv y_2$$

$$y(\bar{\beta}, \underline{a}, \underline{\theta}) = y(\underline{\beta}, \bar{a}, \underline{\theta}) = y(\underline{\beta}, \underline{a}, \bar{\theta}) \equiv y_3$$

$$y(\underline{\beta}, \underline{a}, \underline{\theta}) \equiv y_4.$$

The intricacies of the problem arise because of the pooling possibilities manifest in (1). When y_2 or y_3 is produced, the principal cannot distinguish the effort level of the agent, the productivity of the job, or whether the innovation has been adopted.

The agent derives disutility both from supplying effort and adopting the innovation. We can take \underline{a} , \bar{a} , $\underline{\theta}$, and $\bar{\theta}$ to be real numbers with $\underline{a} < \bar{a}$ and $0 = \underline{\theta} < \bar{\theta}$ and then let the agent's disutility be given by

Action	Disutility
$\bar{a}, \bar{\theta}$	$\bar{a} + \bar{\theta}$
$\underline{a}, \bar{\theta}$	$\underline{a} + \bar{\theta}$
$\bar{a}, \underline{\theta}$	\bar{a}
$\underline{a}, \underline{\theta}$	\underline{a} .

We then assume that $\bar{\theta} > \bar{a} - \underline{a}$, so that $\bar{a} + \bar{\theta} > \underline{a} + \bar{\theta} > \bar{a} > \underline{a}$. This reveals that it is most costly (in utility terms) to supply high effort and adopt the innovation and least costly to do neither. Of the two possible intermediate choices, adopting the innovation with low effort is more costly than supplying high effort without adopting. The principal must provide the agent with nonnegative utility, since the agent retains the option of nonparticipation and receiving a utility which we normalize to zero. In the Soviet Union, for example, managers retain the option of becoming workers (though see footnote 3).

We assume that

$$(2) \quad y_1 > y_2 > y_3 > y_4 \geq \underline{a}.$$

The first three inequalities in condition (2) indicate that output is higher in a high-productivity job than in a low-productivity job and that output can be increased by exerting high rather than low effort and by innovation adoption. The final inequality in (2) ensures that it is always profitable to employ an agent, even if the job is low-productivity and low effort is supplied with no innovation adoption. We further assume that

$$(3) \quad \bar{\theta} > (y_i - y_{i+1}) > (\bar{a} - \underline{a}) \quad i = 1, 2, 3$$

$$(4) \quad \bar{\theta} < 2(\bar{a} - \underline{a}).$$

The first inequality in (3) indicates that the disutility of adopting the innovation exceeds the single-period gain in output. The second inequality indicates that it is profitable to induce the agents to supply high effort and ensures that high effort will be induced in at least some cases. Condition (4) implies that over the course of two periods it is less costly to raise output via innovation adoption than through high effort. Conditions (3) and (4) yield $\bar{\theta} < 2(y_i - y_{i+1})$, which ensures that innovation adoption is efficient in a two-period model.

C. Strategies and Equilibrium

We can now describe formally the strategy spaces and payoffs of the principal and agent and state equilibrium conditions. Let $P_2(\bar{\beta}, \bar{\theta})$ be the principal's period-two expectation that $\beta = \bar{\beta}$ and $\theta = \bar{\theta}$, with $P_2(\bar{\beta}, \underline{\theta})$, $P_2(\underline{\beta}, \bar{\theta})$, and $P_2(\underline{\beta}, \underline{\theta})$ being similar. Let $P_2 \in \{P \in \mathbb{R}_+^4 : \sum P_i = 1\} \equiv S^4$ be a vector of such probabilities (S^4 is the unit simplex in \mathbb{R}^4). The principal provides a remuneration scheme in each period to the agent which specifies the payment, denoted $h_{\cdot i}$, to be made to the agent in the event that the commonly observed outcome in period \cdot is y_i ($i = 1, 2, 3, 4$). The principal's pure strategy set thus consists of triples $h \equiv (h_1, t, h_2)$ where $h_1 \in \mathbb{R}_+^4$, $t \in \{\text{Yes, No}\}$, and $h_2 : S^4 \rightarrow \mathbb{R}_+^4$. The payment attached in period one to output y_i is specified as h_{1i} ; t identifies whether job transfers occur; and

librium strategies but leaves the basic character of the results unchanged.

$h_{2i}(P_2)$ identifies the period-two payment attached to output y_i given expectations P_2 .

Let H_2 be the set of functions $h_2: S^4 \rightarrow \mathbb{R}_+^4$ and let H be the set of triples (h_1, t, h_2) . Then, an agent's strategy is a pair $z_1: \{\beta, \bar{\beta}\} \times H \rightarrow \{0, (\underline{a}, \underline{\theta}), (\underline{a}, \bar{\theta}), (\bar{a}, \underline{\theta}), (\bar{a}, \bar{\theta})\}$ and $z_2: \{\beta, \bar{\beta}\} \times \{\underline{\theta}, \bar{\theta}\} \times H_2 \times S^4 \rightarrow \{0, \underline{a}, \bar{a}\}$: $z_1(\beta, h)$ gives the agent's period-one choice of a [denoted $z_{1a}(\beta, h)$] and θ [denoted $z_{1\theta}(\beta, h)$] as a function of the productivity of the job occupied by the agent and the remuneration scheme chosen by the principal; $z_2(\beta, \theta, h_2, P_2)$ gives the agent's period-two choice of effort as a function of the productivity and innovation status of the job, the period-two remuneration scheme, and the principal's period-two expectation. A choice of zero is taken to denote nonparticipation, yielding a utility and an output of zero.

If output y_{ij} appears in period i , then the principal's payoff in that period is given by $y_{ij} - h_{ij}$. Let $\alpha: \{\text{Yes}, \text{No}\} \rightarrow [1, \alpha]$, denoted $\alpha(t)$, be a function with $\alpha(\text{Yes}) = 1$ and $\alpha(\text{No}) = \alpha$, where $\alpha > 1$ is the productivity parameter capturing the accumulation of job-specific human capital. Then the principal's objective for the game is¹²

$$(5) \max_{h \in H} E_{\beta} \{ y(\beta, z_1(\beta, h)) - h_1(y_1(\beta, z_1(\beta, h))) + \alpha(t) y(\beta, z_2(\beta, z_{1\theta}(\beta, h), h_2, P_2), z_{1\theta}(\beta, h)) - h_2(y(\beta, z_2(\beta, z_{1\theta}(\beta, h), h_2, P_2), z_{1\theta}(\beta, h)), P_2) \}.$$

In the second period, the principal solves

$$(6) \max_{h_2 \in H_2} E_{\beta} \{ \alpha(t) y(\beta, z_2(\beta, z_{1\theta}(\beta, h), h_2, P_2), z_{1\theta}(\beta, h)) - h_2(y(\beta, z_2(\beta, z_{1\theta}(\beta, h), h_2, P_2), z_{1\theta}(\beta, h)), P_2) \}.$$

TABLE 4—AGENT'S PERIOD-ONE UTILITY FROM VARYING CHOICES

Agent's action	Job	Outcome	Utility
$\bar{a}, \bar{\theta}$	$\bar{\beta}$	y_1	$h_{11} - (\bar{a} + \bar{\theta})$
$\bar{a}, \underline{\theta}$	$\bar{\beta}$	y_2	$h_{12} - \bar{a}$
$\underline{a}, \bar{\theta}$	$\bar{\beta}$	y_2	$h_{12} - (\underline{a} + \bar{\theta})$
$\underline{a}, \underline{\theta}$	$\bar{\beta}$	y_3	$h_{13} - \underline{a}$
$\bar{a}, \bar{\theta}$	$\underline{\beta}$	y_2	$h_{12} - (\bar{a} + \bar{\theta})$
$\bar{a}, \underline{\theta}$	$\underline{\beta}$	y_3	$h_{13} - \bar{a}$
$\underline{a}, \bar{\theta}$	$\underline{\beta}$	y_3	$h_{13} - (\underline{a} + \bar{\theta})$
$\underline{a}, \underline{\theta}$	$\underline{\beta}$	y_4	$h_{14} - \underline{a}$

The agent's payoff in period i is given by $h_i(y_i)$ minus the disutility of the period- i effort and innovation adoption choice. The agent's single-period utility levels from varying choices are given in Table 4. In the second period, the agent in a β job (for $\beta = \underline{\beta}$ or $\bar{\beta}$) thus solves

$$(7) \max_{z_2 \in \{\underline{a}, \bar{a}\}} \{ h_2(y(\beta, z_2, z_{1\theta}(\beta, h)), P_2) - z_2 \}.$$

In period one, the agent in a β ($= \underline{\beta}$ or $\bar{\beta}$) job solves

$$(8) \max_{\substack{z_{1a} \in \{\underline{a}, \bar{a}\} \\ z_{1\theta} \in \{\underline{\theta}, \bar{\theta}\} \\ z_2 \in \{\underline{a}, \bar{a}\}}} \{ [h_1(y(\beta, z_{1a}, z_{1\theta})) - z_{1a} - z_{1\theta}] + [h_2(y(\beta, z_2, z_{1\theta}), P_2) - z_2] \}.$$

A subgame perfect equilibrium is then a triple of strategies (h, z_1, z_2) such that the h_2 component of h maximizes (6) given z_2 ; h maximizes (5) given z_1 and z_2 and given that h_2 must solve (6); $z_2(\beta)$ and $z_2(\bar{\beta})$ each maximizes (8) given h ; $z_1(\beta)$ and $z_1(\bar{\beta})$ each maximizes (7) given h and given that $z_2(\beta)$ and $z_2(\bar{\beta})$ solve (8); and period-two posterior expectations P_2 are calculated according to Bayes' rule. This last requirement warrants some elaboration. If job transfers are not practiced, this requires that the P_2 appearing in (5)–(8) be calculated by the principal according to

¹²In keeping with our consideration of a single agent, payoffs or profits for the principal will always be taken to mean expected profits per agent.

Bayes' rule and that both the principal and agent recognize that the agent's actions, through their effects on y_{1i} , affect P_2 . If job transfers are practiced, then the principal again calculates P_2 via Bayes' rule and recognizes that the actions the agent is induced to take in period one will affect P_2 . The agent's maximization takes P_2 to be equal to the equilibrium value calculated by the principal but to be exogenously fixed at that level (because P_2 applies to a different period-two job than the one the agent now occupies).

III. Equilibrium Existence and Characterization

This section presents a basic equilibrium existence and characterization result:

PROPOSITION 1: *A subgame perfect equilibrium exists. Generically,*

- 1) *the equilibrium is unique;*
- 2) *the equilibrium strategies are pure;*
- 3) *the period-one outcome is separating, in that period-one outputs reveal job productivities;*
- 4) *all agents are induced to supply high effort in period two;*
- 5) *the equilibrium may or may not involve job transfers, depending upon parameter values;*
- 6) *innovation and high effort are induced from agents in β jobs in period one;*
- 7) *depending upon parameter values, agents in β jobs may be induced to supply high effort and innovate (this may occur with or without job transfers), may be induced to supply high effort and not innovate (only with job transfers), or may be induced to supply low effort and not innovate (only without job transfers);*
- 8) *the equilibrium will consist of one (and only one) of four sets of strategies, each corresponding to one of the four possibilities identified in 7, depending upon parameter values. The four sets of strategies are given in Table 5.¹³ The payoff to the princi-*

pal from each potential equilibrium is given in Table 6.

The characteristics of the four potential equilibria are listed in Table 6.

The Appendix proves this proposition. The intuition behind these results is readily provided. First, the existence and uniqueness of the equilibrium follows from a backward induction argument. Second, the equilibrium features pure strategies, because the principal is in general not indifferent over agents' actions. An equilibrium then cannot exhibit agent randomization, because the principal would respond by slightly increasing the payoff to the principal's preferred outcome and hence inducing pure strategies.¹⁴ Third, the principal induces separating outcomes in period one, because the information gleaned from such outcomes allows the principal to reduce the cost of period-two contracts. Fourth, given such separation, there are no information-based obstacles to inducing effort choices in period two, and it is profitable for the principal to induce all agents to expend high effort in the second period. Fifth, the period-one separating outcome may or may not be achieved with the help of job transfers, depending upon the values of parameters that determine the rate at which transfers trade reduced period-one incentive costs for sacrifices of human capital accumulation. Sixth, agents in high-productivity jobs will always be induced to adopt and supply high effort, because the output gains from doing so exceed the utility costs (and

¹³In order to avoid clutter, Table 5 presents only the actions that agents' strategies yield along the equilibrium path. Out-of-equilibrium behavior is easily calculated from (7) and (8).

¹⁴We are assuming here and throughout the analysis that when an agent is indifferent between two actions, the agent chooses the action most preferred by the principal. This presumption is required for the existence of an equilibrium. If the agent does not choose the principal's most preferred action when the agent is indifferent, the principal could induce such a choice by adding an arbitrarily small ϵ to the payoff from the desired action. As there is no smallest such ϵ which will suffice, equilibrium requires $\epsilon = 0$ and a tie which is broken in the principal's favor. It is important to note that it is not entirely obvious that ties will be broken in the principal's favor, especially with multiple agents. Ching-to Ma (1987, 1988) and Ma et al. (1988) examine incentive schemes that do not invoke such an assumption.

TABLE 5—EQUILIBRIUM STRATEGIES

Equilibrium	Strategy	Period 1	Period 2		Transfer?	
			$P_2 = 1$	$P_2 < 1$		
1	Principal:	h_1	$2\bar{\theta} + \bar{a}$	\bar{a}	—	yes
		h_2	$\bar{\theta} + \bar{a}$	\underline{a}	\bar{a}	
		h_3	\underline{a}	—	\underline{a}	
		h_4	\underline{a}	—	—	
	Agent:	$\bar{\beta}$	$\bar{a}, \bar{\theta}$	\bar{a}		
		$\underline{\beta}$	$\bar{a}, \bar{\theta}$		\bar{a}	
2	Principal:	h_1	$\bar{\theta} + 2\bar{a} - \underline{a}$	\bar{a}	—	yes
		h_2	\bar{a}	\underline{a}	—	
		h_3	\bar{a}	—	\bar{a}	
		h_4	\underline{a}	—	\underline{a}	
	Agent:	$\bar{\beta}$	$\bar{a}, \bar{\theta}$	\bar{a}		
		$\underline{\beta}$	$\bar{a}, \underline{\theta}$		\bar{a}	
3	Principal:	h_1	$\bar{\theta} + 3\bar{a} - 2\underline{a}$	\bar{a}	—	no
		h_2	$\bar{\theta} + \bar{a}$	\underline{a}	\bar{a}	
		h_3	\underline{a}	—	\underline{a}	
		h_4	\underline{a}	—	—	
	Agent:	$\bar{\beta}$	$\bar{a}, \bar{\theta}$	\bar{a}		
		$\underline{\beta}$	$\bar{a}, \bar{\theta}$		\underline{a}	
4	Principal:	h_1	$\bar{\theta} + \bar{a}$	\bar{a}	—	no
		h_2		\underline{a}	—	
		h_3	—	—	\bar{a}	
		h_4	\underline{a}	—	\underline{a}	
	Agent:	$\bar{\beta}$	$\bar{a}, \bar{\theta}$	\bar{a}		
		$\underline{\beta}$	$\underline{a}, \underline{\theta}$		\underline{a}	

Note: P_2 is $P_2(\bar{\beta}, \bar{\theta})$.

TABLE 6—SUMMARY OF POTENTIAL EQUILIBRIUM OUTCOMES AND PRINCIPAL'S EXPECTED PROFITS

Equilibrium	Period-1 actions (a, θ) and outcome (y) in $\bar{\beta}$ job	Period-1 actions (a, θ) and outcome (y) in $\underline{\beta}$ job	Transfers practiced?	$\underline{\beta}$ adopt innovation?	$\underline{\beta}$ supply high effort?
1	$\bar{a} \bar{\theta} y_1$	$\bar{a} \bar{\theta} y_2$	yes	yes	yes
2	$\bar{a} \bar{\theta} y_1$	$\bar{a} \bar{\theta} y_3$	yes	no	yes
3	$\bar{a} \bar{\theta} y_1$	$\bar{a} \bar{\theta} y_2$	no	yes	yes
4	$\bar{a} \bar{\theta} y_1$	$\underline{a} \underline{\theta} y_4$	no	no	no

Principal's expected profits:

$$\begin{aligned}\pi_1 &= p_1[2y_1 - 2\bar{\theta} - 2\bar{a}] + (1 - p_1)[2y_2 - \bar{\theta} - 2\bar{a}] \\ \pi_2 &= p_1[2y_1 - \bar{\theta} - 3\bar{a} + \underline{a}] + (1 - p_1)[2y_3 - 2\bar{a}] \\ \pi_3 &= p_1[(1 + \alpha)y_1 - 4\bar{a} - \bar{\theta} + 2\underline{a}] + (1 - p_1)[(1 + \alpha)y_2 - 2\bar{a} - \bar{\theta}] \\ \pi_4 &= p_1[(1 + \alpha)y_1 - 2\bar{a} - \bar{\theta}] + (1 - p_1)[y_4 + \alpha y_3 - \bar{a} - \underline{a}]\end{aligned}$$

inducing these agents to do so does not raise the incentive costs associated with other agents). Seventh, agents in low-productivity jobs may or may not be induced to innovate, depending upon the values of parameters that determine the relative magnitudes of the increased output and the increase in incentive costs associated with high-productivity agents. Finally, we then have four possible equilibria, differing according to whether transfers are practiced and whether low-productivity agents are induced to innovate.

IV. Incentives and Innovation

The four sets of strategies given in Table 5 represent four possible equilibria. Because the principal moves first in a sequential game, we can equivalently view these as four possible optimal remuneration schemes for the principal (with the associated induced-agent actions). We begin with the question of whether there exist parameter values for which it is optimal for the principal to offer each remuneration scheme.

COROLLARY 1: *There exist parameter values for which each of equilibria 1–4 is the unique equilibrium.*

PROOF:

We prove this by presenting four examples. In each example,

$$y_1 = 26 \quad p_1 = 0.3$$

$$y_2 = 16.5 \quad \bar{a} = 7.8$$

$$y_3 = 10.6 \quad \underline{a} = 2$$

$$y_4 = 2.$$

Example 1—Remuneration scheme 1 optimal:

Parameters	Payoffs		
$\bar{\theta} = 9.6$	$\pi_1 = 10.62$	$\pi_3 = 10.2135$	
$\alpha = 1.01$	$\pi_2 = 10.22$	$\pi_4 = 10.1522.$	

Example 2—Remuneration scheme 2 optimal:

Parameters	Payoffs		
$\bar{\theta} = 11.7$	$\pi_1 = 7.89$	$\pi_3 = 8.1135$	
$\alpha = 1.01$	$\pi_2 = 9.59$	$\pi_4 = 9.522.$	

Example 3—Remuneration scheme 3 optimal:

Parameters	Payoffs		
$\bar{\theta} = 9.6$	$\pi_1 = 10.62$	$\pi_3 = 11.955$	
$\alpha = 1.1$	$\pi_2 = 10.22$	$\pi_4 = 11.522.$	

Example 4—Remuneration scheme 4 optimal:

Parameters	Payoffs		
$\bar{\theta} = 11.7$	$\pi_1 = 7.89$	$\pi_3 = 9.855$	
$\alpha = 1.1$	$\pi_2 = 9.59$	$\pi_4 = 10.892.$	

The potential optimality of remuneration schemes 2 and 4 confirms the argument presented in the introduction that it may be unprofitable to induce the low-productivity agent to adopt the innovation, because of the increased incentive costs of inducing the high-productivity agent to innovate. Notice that transfers occur when α is low (job-specific human capital is not important) and that innovation is induced from all agents when $\bar{\theta}$ is low (innovation is inexpensive).

We can pursue the connection between parameter values and the optimal remuneration scheme further. First, we investigate the conditions under which the principal will find it optimal to induce innovation adoption from all agents. Comparing schemes 1 and 2, we find that the former entails an extra cost of $\bar{\theta} - p_1(\bar{a} - \underline{a})$ in exchange for an expected output gain of $2(1 - p_1)(y_2 - y_3)$. Similarly, comparing schemes 3 and 4 reveals that the former entails an extra cost of $(1 - p_1)\bar{\theta} + p_1[2(\bar{a} - \underline{a})] + (1 - p_1)(\bar{a} - \underline{a})$ in exchange for an extra output

gain of $(1 - p_1)[(1 + \alpha)y_2 - \alpha y_3 - y_4]$. We thus immediately have the following corollary (we can identify precise boundaries on the parameter values for which innovation adoption will occur, but the resulting expressions are cumbersome¹⁵).

COROLLARY 2: *Inducing all agents to adopt the innovation is more likely to be optimal as $\bar{\theta}$ is small, $y_2 - y_3$ large, α large, $y_2 - y_4$ large, and p_1 small.*

These results are not surprising. They indicate that inducing innovation adoption from low-productivity agents is more likely to be optimal when the cost of innovation ($\bar{\theta}$) is small, the output gains ($\alpha[y_2 - y_3]$ and $y_3 - y_4$) are large, and it is more likely that jobs are low-productivity (p_1 is small).

The effect of variations in the marginal cost of inducing high effort, $(\bar{a} - \underline{a})$, on the optimality of inducing innovation adoption from low-productivity agents is ambiguous. If job transfers are not practiced, so that schemes 3 and 4 are relevant, increases in $(\bar{a} - \underline{a})$ make it less likely that low-productivity-agent innovation adoption is optimal. This occurs because, without job transfers, agents in β jobs are induced to adopt the innovation if and only if they also are induced to supply high effort, so that an increase in the cost of the latter makes it less

likely that inducing innovation adoption is optimal. In contrast, if job transfers are practiced, so that remuneration schemes 1 and 2 are relevant, then increases in $(\bar{a} - \underline{a})$ make it more likely that innovation adoption by agents in low-productivity jobs is optimal. This occurs because job transfers reduce the cost of inducing high effort. In the presence of job transfers, agents in the β job are then induced to supply high effort regardless of whether they adopt the innovation, and increases in the cost of inducing high effort are not inimical to inducing innovation adoption.

These findings direct attention to the role in the model played by the agents' effort choices. One might initially wonder why we do not strip the model of effort choices and concentrate on the choice of innovation adoption in jobs of varying productivity. Effort choices are essential to the analysis because the most profitable deviation for an agent from a recommendation of high effort and innovation adoption is to adopt the innovation but supply low effort. Adopting the innovation raises the productivity of the job while supplying low effort saves on effort disutility and (most importantly) disguises the job's productivity. This allows the agent to avoid more severe future remuneration schemes while making it easier to attain high payments from existing schemes. The highest incentive costs accordingly arise in deterring agents from adopting the innovation while supplying low effort, and an important facet of the incentive problem is not captured when effort choices are ignored.

Attention now turns to job transfers. Ickes and Samuelson (1987) demonstrate that job transfers may reduce the cost of effort incentives. This is again evident in the results of this paper. A comparison of remuneration schemes 1 and 3, for example, reveals that an extra ratchet price of $2(\bar{a} - \underline{a}) - \bar{\theta}$ is required to induce high effort from agents in β jobs in scheme 3 (without job transfers), which is unnecessary in scheme 1 (with transfers). Because of this, job transfers are optimal in some circumstances. In addition, if an agent in a β job is not induced to adopt the innovation, as in schemes 2 and 4,

¹⁵For example, inducing innovation adoption from low-productivity agents is optimal if $\pi_1 > \max\{\pi_2, \pi_4\}$ or $\pi_3 > \max\{\pi_2, \pi_4\}$. After manipulation, this is equivalent to

$$\begin{aligned} \bar{\theta} &< \max \left[\min \left(p_1[\bar{a} - \underline{a}] + (1 - p_1)[2(y_2 - y_3)], \right. \right. \\ &\quad p_1[(1 - \alpha)y_1] + (1 - p_1) \\ &\quad \times [2y_3 - y_4 - \alpha y_3 - (\bar{a} - \underline{a})]), \\ \min &\left(\frac{p_1}{1 - p_1} [(\alpha - 1)y_1 - (\bar{a} - \underline{a})] + [(1 + \alpha)y_2 - 2y_3], \right. \\ &\quad \frac{p_1}{1 - p_1} [-2(\bar{a} - \underline{a})] \\ &\quad \left. \left. + [(1 + \alpha)y_2 - \alpha y_3 - y_4 - (\bar{a} - \underline{a})] \right) \right]. \end{aligned}$$

high effort is optimally induced from this agent only if job transfers are practiced. This occurs because the cost of such effort is prohibitive (given $\underline{\beta}$ does not adopt) without job transfers.

We can identify conditions under which job transfers will occur as follows.

COROLLARY 3: *Job transfers are more likely to be optimal as α is small, p_1 is close to neither 0 nor 1, and $y_2 - y_3$ and $y_3 - y_4$ are large.*

This follows immediately from comparing profit expressions for job-transfer schemes 1 and 2 with those for remuneration schemes 3 and 4. The results are expected. First, job transfers sacrifice job-specific human capital and, thus, are most likely to be optimal when the effect of such capital, or α , is small. Second, job transfers are designed to deter agents from concealing job productivities, a problem that is most serious when the principal entertains significant uncertainty concerning productivity, so that p_1 is close to neither 0 nor 1. Finally, job transfers are most likely to be optimal when the effect of high effort, given by $y_2 - y_3$ and $y_3 - y_4$, is large. Notice that the effect of effort costs is again ambiguous. As $(\bar{a} - \underline{a})$ increases, transfers are more likely to be optimal if all agents are induced to adopt the innovation but less likely to be optimal if $\underline{\beta}$ agents are not induced to adopt.

We can now make precise the statement that, in the hierarchical system, the cost of innovation adoption increases at an increasing rate. The principal in our model has the option of offering a remuneration scheme that induces the adoption of innovation from no agents, from agents in high-productivity jobs only, or from agents in all jobs. Let C_0 be the cost of inducing innovation in no jobs, C_1 the cost of inducing innovation in high-productivity jobs only, and C_2 the cost of inducing innovation from all agents.

PROPOSITION 2: *Suppose job transfers are not practiced. Then for all $p_1 \in [0, 1)$,*

$$(9) \quad C_2 - C_1 > C_1 - C_0.$$

Alternatively, if job transfers are practiced, then (9) holds for all $p_1 \in [0, \bar{\theta}/[\bar{\theta} + (\bar{a} - \underline{a})]]$.

PROOF:

Consider the case of job transfers. We have

$$C_0 = [p_1(2\bar{a} - \underline{a}) + (1 - p_1)\bar{a}] + \bar{a}$$

$$C_1 = [p_1(\bar{\theta} + 2\bar{a} - \underline{a}) + (1 - p_1)\bar{a}] + \bar{a}$$

$$C_2 = [p_1(2\bar{\theta} + \bar{a}) - (1 - p_1)(\bar{\theta} + \bar{a})] + \bar{a}.$$

The bracketed term in each case gives the period-one cost of inducing the outcome, which is the sum of the remunerations received by a $\bar{\beta}$ and $\underline{\beta}$ agent multiplied by p_1 and $1 - p_1$, respectively. The second term is the period-two cost. Because these are separating outcomes, both agents receive \bar{a} in period two, and the period-two cost is thus $p_1\bar{a} + (1 - p_1)\bar{a} = \bar{a}$. The terms C_1 and C_2 follow directly from Table 6, while C_0 follows from Ickes and Samuelson (1987). We then calculate $C_2 - C_1 > C_1 - C_0$ if $p_1 \in [0, \bar{\theta}/[\bar{\theta} + (\bar{a} - \underline{a})]]$. The no-transfers case involves a similar calculation.

The result shows that the cost of inducing agents in both types of jobs to adopt the innovation is more than twice that of inducing agents in only one type of job to adopt the innovation. This is what we refer to as the cost of innovation adoption increasing at an increasing rate. A first expectation is that this result will hold only if p_1 is not too large. If p_1 is large, there are many more $\bar{\beta}$ jobs than $\underline{\beta}$ jobs, and the incremental cost of inducing the relatively few agents in $\underline{\beta}$ jobs to adopt the innovation would appear unlikely to be larger than the incremental cost of inducing the relatively many agents in $\bar{\beta}$ jobs to adopt. Without transfers, however, we find $C_2 - C_1 > C_1 - C_0$ for all $p_1 < 1$. This result appears because agents in $\underline{\beta}$ jobs can be induced to adopt only if incentive costs are also paid to agents in $\bar{\beta}$ jobs. The incremental cost of inducing agents in $\underline{\beta}$ jobs to adopt then exceeds the corresponding adoption cost for agents in $\bar{\beta}$ jobs,

regardless of the relative numbers of each type of job. Because job transfers partially alleviate the incentive problem (though at the cost of sacrificing job-specific human capital), an upper bound on p_1 (which always exceeds $1/2$) is required for the comparison with job transfers.

V. Conclusion

We have examined the incentives to adopt innovations in a hierarchical system such as the Soviet planned-enterprise sector. Because the principal in such a system is generally uncertain as to the productivity of the jobs filled by the agents, a ratchet effect appears. An agent's exemplary performance is taken as a signal that the agent fills a high-productivity job and is accordingly followed by more demanding remuneration schemes. Agents then have an incentive to disguise the productivity of their jobs.

The operation of the ratchet effect raises particularly severe problems in inducing the adoption of innovations. The principal will always induce innovation adoption from agents in high-productivity jobs and will do so by attaching a large payment to the relatively high output accompanying innovation. Suppose now that agents in lower-productivity jobs are to be induced to adopt an innovation. To do so, the payments attached to the outputs produced if these agents innovate must be increased. Unfortunately, this increases the return that agents in higher-productivity jobs can earn by deviating from recommended actions in order to disguise the productivities of their jobs. The payments made to these agents must then also be increased to deter such deviations. As a result, each decision by the principal to induce innovation adoption from agents in jobs of a given productivity level increases the incentive costs of inducing innovation adoption from all agents in jobs of higher productivity. The principal will thus begin by inducing innovation adoption in the highest-productivity jobs and proceed downward, finding that at each step the costs of inducing innovation adoption increase at an increasing rate. The response

to these incentive costs is likely to entail inducing a relatively low rate of innovation adoption.

We have derived these results in a highly stylized model, and we can comment on which features of the model are most important. Similar forces will appear if there are more than two levels of θ , a , and α , though the analysis is more complicated (and completely separating contracts are unlikely to be optimal or even possible if the number of values is too large or forms a continuum). The basic forces also survive generalization to many or infinite periods, though this generalization appears to be most interesting if job productivities are continually subjected to random shocks so that there is always new information to be learned. The simplicity of the results, especially those pertaining to job transfers, is driven by the separation of the adverse selection and moral-hazard problems, with the former applying only to jobs and the latter only to agents' actions. The principal faces a much more difficult problem if agents also differ in types. In some cases, however, job transfers will still be a useful device to reduce the incentive to disguise job productivities, though transfers will be ineffective in eliminating incentives to disguise worker productivities.

We can use these results to illustrate the key difference in inducing innovation adoption between a centralized system and a decentralized or market economy. Suppose that output y sells at a fixed price (which we can normalize to equal unity). Then an innovation will be adopted whenever the increment to output, or $2(y_1 - y_{i+1})$, exceeds the direct cost of adopting, or θ . The market system would thus always induce the adoption of the innovations examined in our model. This yields an efficient level of investment and allows a Pareto efficient outcome to appear. Equivalently, we can say that the curve identifying the private cost of adopting an innovation in the various enterprises in the market economy is linear in the number of adoptions and has a constant slope or marginal cost equal to the technical cost of the innovation. Any innovation whose private benefits exceed this

technical cost is then adopted. In addition, the market economy eliminates the incentive-cost externalities examined above, so that private and social costs and benefits coincide, yielding an efficient innovation level.

In a hierarchical system, in contrast, the corresponding cost curve increases at an increasing rate. The marginal cost of adopting an innovation equals the technical cost of adoption only for initial adoptions. The marginal cost of subsequent adoptions includes increasingly large increases in incentive costs. Given this more sharply increasing cost curve, the optimal response is the inducement of fewer innovation adoptions than in the decentralized economy. In particular, cases will arise in which innovations exist whose benefits exceed the direct cost of adoption but which are not adopted. This yields an outcome with inefficiently low innovation adoption. It appears as if this difficulty is inherent in the hierarchical nature of the system. Mere adjustments in incentive schemes within the hierarchical system are unlikely to counter the problem.

Finally, we can return to the case of the Soviet Union. We have seen that the optimal response to the increasing cost-of-adoption schedule that arises in a hierarchical system is relatively little innovation. While the Soviet Union exhibits all of the characteristics of a hierarchical system required to yield the increasing cost-of-adoption curve and also exhibits relatively little innovation, it is clear that the Soviets do not consider their innovation adoption rates to be optimal. Comments such as those of Malenkov and Gorbachev, quoted in our introduction, suggest frustration with achieved performance. Reinforcing this, Gorbachev has designated "the acceleration of scientific and technical progress" to be "problem number one" for the USSR (see Amman and Cooper, 1986 p. 1). On the one hand, this perceived lack of optimality reflects the lack of coordination in Soviet planning and a resulting inability to extract the hierarchical system's best performance. In our view, however, these sentiments also represent a frustration with the constraints imposed by

the hierarchical system and a desire to be able to operate without such constraints. This is reflected in the comment by Abel Aganbegyan, one of Gorbachev's chief economic advisers, that current trends can be overcome only by "revolutionary changes" (Aganbegyan, 1988, p. 84) and by the reforms which form the heart of *perestroika*. Our findings suggest that achieving increased adoption rates without prohibitive cost may require not just a tinkering with the form of incentive contracts but a modification of the hierarchical decision-making process, so that *perestroika* faces a formidable task.

APPENDIX

This Appendix proves Proposition 1 via a series of lemmas. In many cases, the proofs are straightforward adaptations of previous proofs or arguments appearing in Ickes and Samuelson (1987) and are omitted. Full details are available in Dearden et al. (1989).

LEMMA 1: *In equilibrium, an agent in a β job earns a zero payoff in period two. If the period-one outcome is separating, so that one of $P_2(\bar{\beta}, \bar{\theta})$, $P_2(\bar{\beta}, \underline{\theta})$, $P_2(\underline{\beta}, \bar{\theta})$, or $P_2(\underline{\beta}, \underline{\theta})$ equals unity, then the corresponding agent earns a zero payoff in period two.*

PROOF:

This follows directly from the period-two equilibrium conditions given by (6) and (7). In particular, the principal will find it optimal to reduce the payments h_{2i} , $i = 1, \dots, 4$, until some agent earns a return of zero. The only question concerns what type of job that agent occupies. If one of $P_2(\bar{\beta}, \bar{\theta})$, $P_2(\bar{\beta}, \underline{\theta})$, $P_2(\underline{\beta}, \bar{\theta})$, or $P_2(\underline{\beta}, \underline{\theta})$ equals unity, then the utility of an agent in the corresponding job will be reduced to zero, giving the second result of Lemma 1. Suppose there is positive probability of either a β or a $\bar{\beta}$ job. Because the agent in a $\bar{\beta}$ job can always produce as much output and hence secure as much utility as an agent in a β job, the latter's utility will then be the one reduced to zero, giving the first statement of Lemma 1.

LEMMA 2: *In equilibrium, the $\bar{\beta}$ agent chooses $(\bar{a}, \bar{\theta})$ in period one with probability one.*

PROOF:

Suppose not. We will show that h_1 can be adjusted so as to increase the principal's profits. If $(\bar{a}, \bar{\theta})$ is played with positive probability, say ρ , the principal can increase h_{11} to $h_{11} + \varepsilon$ for small ε . This breaks the $\bar{\beta}$ agent's indifference, causing the agent to play $(\bar{a}, \bar{\theta})$ with unitary probability. This gives an output gain of at least $p_1(1 - \rho)(y_1 - y_2)$ at a cost of ε , which is profit-increasing for the principal for sufficiently small ε . Suppose then that $(\bar{a}, \bar{\theta})$ is played with zero probability and that $(\underline{a}, \underline{\theta})$, $(\bar{a}, \underline{\theta})$, and $(\underline{a}, \underline{\theta})$ are played with probabilities ρ_A , ρ_B , and ρ_C , each of which is positive (the extension to the case in which one or more of these equals zero is immediate). Then the indifference needed to support this randomization requires $h_{12} - \underline{a} - \bar{\theta} + x_A = h_{12} - \bar{a} + x_B = h_{13} - \underline{a} + x_C$, where x_A is the expected period-two return to the $\bar{\beta}$ agent given that $(\underline{a}, \underline{\theta})$ is played in period one and x_B and x_C are analogous for $(\bar{a}, \underline{\theta})$ and $(\underline{a}, \underline{\theta})$. Now set h_{11} so that the $\bar{\beta}$ agent is indifferent between $(\bar{a}, \bar{\theta})$ and $(\underline{a}, \underline{\theta})$, or $(\bar{a}, \underline{\theta})$. This requires $h_{11} - \bar{a} - \bar{\theta} = h_{12} - \underline{a} - \bar{\theta} + x_A = h_{12} - \bar{a} + x_B = h_{13} - \underline{a} + x_C$ or

$$(A1) \quad h_{11} - h_{12} = \bar{a} - \underline{a} + x_A$$

$$h_{11} - h_{12} = \bar{\theta} + x_B$$

$$h_{11} - h_{13} = (\bar{a} - \underline{a}) + \bar{\theta} + x_C.$$

This choice of h_{11} induces the $\bar{\beta}$ agent to choose $(\bar{a}, \bar{\theta})$ with probability one (otherwise add ε to h_{11}). The cost to the principal of setting this value of h_{11} , from (A1) is then at most

$$(A2) \quad p_1[\rho_A(\bar{a} - \underline{a} + x_A) + \rho_B(\bar{\theta} + x_B) + \rho_C(\bar{a} + \bar{\theta} - \underline{a} + x_C)]$$

where p_1 is the probability of a $\bar{\beta}$ agent.¹⁶ The gain to the principal from setting this value of h_{11} is at least

$$(A3) \quad p_1\{\rho_A[y_1 - y_2 + x_A] + \rho_B[2(y_1 - y_2) + x_B] + \rho_C[y_1 - y_3 + y_1 - y_2 + x_C]\}.$$

From (3) and (4), we now see that (A3) exceeds (A2). This precludes the optimality of an outcome in which the $\bar{\beta}$ agent mixes over any of $(\underline{a}, \underline{\theta})$, $(\bar{a}, \underline{\theta})$, and $(\underline{a}, \underline{\theta})$ and completes the proof.

LEMMA 3: *The β agent does not play $(\underline{a}, \bar{\theta})$ with positive probability in period one.*

PROOF:

The β agent strictly prefers playing $(\bar{a}, \underline{\theta})$ to $(\underline{a}, \bar{\theta})$, since $(\bar{a}, \underline{\theta})$ provides an identical period-one payment of h_{13} , yields less disutility, and trivially allows the β agent to still reap the equilibrium period-two utility of zero.

LEMMA 4: *The β agent plays a pure strategy in period one.*

PROOF:

Analogous to Lemma 2.

LEMMA 5: *There are five possible equilibrium paths, described in Table A1 where the period-two equilibria are as given in Table A2.*

PROOF:

Lemmas 1–4 indicate that, in equilibrium, agents in $\bar{\beta}$ jobs must play $(\bar{a}, \bar{\theta})$ in period one while those in β jobs must play a pure strategy of either $(\bar{a}, \bar{\theta})$, $(\bar{a}, \underline{\theta})$, or $(\underline{a}, \underline{\theta})$.

¹⁶Since y_1 is infeasible for an agent in a β job, the only implications of this adjustment of h_{11} for an agent in a β job is that it may affect $P_2(\beta, \theta)$ and hence the period-two remuneration scheme. Since the β agent earns zero utility in any period-two remuneration scheme (Lemma 1), this adjustment cannot affect the actions of an agent in a β job and hence does not affect the outcomes or costs that appear if the job is β .

TABLE A1—SUMMARY OF POTENTIAL EQUILIBRIUM PATHS

Outcome path	Period-one actions (a, θ) and outcome (y) in β job	Period-one outcome type: separating (S) or pooling (P)	Period-one actions (a, θ) and outcome (y) in $\underline{\beta}$ job	Job transfers practiced?	Period-two equilibrium
1	$\bar{a} \bar{\theta} y_1$	S	$\bar{a} \bar{\theta} y_2$	yes	1A
2	$\bar{a} \bar{\theta} y_1$	S	$\bar{a} \underline{\theta} y_3$	yes	2A
3	$\bar{a} \bar{\theta} y_1$	S	$\bar{a} \bar{\theta} y_2$	no	1A
4	$\bar{a} \bar{\theta} y_1$	S	$\underline{a} \underline{\theta} y_4$	no	2A
5	$\bar{a} \bar{\theta} y_1$	S	$\bar{a} \underline{\theta} y_3$	no	2A

TABLE A2—PERIOD-TWO EQUILIBRIA

Strategy	Period-two equilibrium 1A ^a	
	$P_2(\bar{\beta}, \bar{\theta}) < p^*$	$P_2(\bar{\beta}, \bar{\theta}) \geq p^*$
Principal: h_{21}	$2\bar{a} - \underline{a}$	\bar{a}
h_{22}	\bar{a}	\underline{a}
h_{23}	\underline{a}	\underline{a}
h_{24}	—	—
Agent: ^b $z_2(\bar{\beta}, \cdot)$	$\bar{a}(y_1)$	$\bar{a}(y_1)$
$z_2(\underline{\beta}, \cdot)$	$\bar{a}(y_2)$	$\underline{a}(y_3)$
Strategy	Period-two equilibrium 2A	
Principal: h_{21}	\bar{a}	
h_{22}	\underline{a}	
h_{23}	\bar{a}	
h_{24}	\underline{a}	
Agent: ^b $z_2(\bar{\beta}, \cdot)$	$\bar{a}(y_1)$	
$z_2(\underline{\beta}, \cdot)$	$\underline{a}(y_4)$	

^aWhere $p^* = 1 - \frac{\bar{a} - \underline{a}}{\alpha(t)(y_2 - y_3)}$.

^bAgent's output is given in parentheses.

Depending upon whether job transfers occur, this yields six period-one outcomes. However, it is suboptimal for the principal to induce $(\bar{a}, \bar{\theta})$ from β agents and $(\underline{a}, \underline{\theta})$ from β agents and to practice transfers. In particular, agents in β jobs would then produce output y_4 , and there would be no opportunity for agents in β jobs to pool. Accordingly, there is no reason to sacrifice job-specific human capital by transferring, and this outcome path is suboptimal. This leaves the five paths listed in Table A1. It remains to show that a unique period-two equilibrium, given by 1A or 2A, can be

associated with each path. This follows the analysis of Ickes and Samuelson (1987).

LEMMA 6: *Only outcome paths 1–4 constitute potential equilibria. The equilibrium associated with each path and the principal's profits are as shown in Tables 5 and 6.*

PROOF:

We examine the first path. The others are analogous. In the first case, the first-period outcome reveals the productivity of a job, with an outcome of y_1 (y_2) indicating that the job has high (low) productivity. The period-two remuneration scheme will be 1A and for any job will induce high effort and yield the agent a period-two utility of zero. The complete remuneration scheme is shown in Table 5.¹⁷ It remains to show that the principal's strategy is optimal (i.e., that it induces the desired outcome at minimum cost). Consider the choices of actions available to the agents in period one and the resulting utilities reported in Table A3. The optimal action for an agent in either a high- or low-productivity job is clearly $(\bar{a}, \bar{\theta})$, as desired. Given the payoffs to the alternative choices of $(\underline{a}, \underline{\theta})$ for an agent in a β job and $(\bar{a}, \bar{\theta})$ for an agent in a β job, the scheme also induces the desired outcomes at mini-

¹⁷A complete specification of a remuneration scheme must also identify the inferences drawn by the principal if an out-of-equilibrium first-period outcome of y_3 or y_4 appears. We assume here that the principal assumes that such an outcome reveals the job to be of low productivity. It is easily verified that this does not disrupt the equilibrium. Similar choices apply to subsequent remuneration schemes, and we omit the details.

TABLE A3—AGENT'S TWO-PERIOD UTILITIES FROM VARYING CHOICES GIVEN REMUNERATION SCHEME 1

Agent	Action	Period-one utility	Period-two utility	Total utility
$\underline{\beta}$	$(\bar{a}, \bar{\theta})$	0	0	0
	$(\underline{a}, \bar{\theta})$	$-\bar{\theta}$	0	$-\bar{\theta}$
	$(\bar{a}, \underline{\theta})$	$\underline{a} - \bar{a}$	0	$-(\bar{a} - \underline{a})$
	$(\underline{a}, \underline{\theta})$	0	0	0
$\bar{\beta}$	$(\bar{a}, \bar{\theta})$	$\bar{\theta}$	0	$\bar{\theta}$
	$(\underline{a}, \bar{\theta})$	$\bar{a} - \underline{a}$	0	$\bar{a} - \underline{a}$
	$(\bar{a}, \underline{\theta})$	$\bar{\theta}$	0	$\bar{\theta}$
	$(\underline{a}, \underline{\theta})$	0	0	0

mum cost. A key step in these calculations is the verification that an agent receives a zero payoff in period two. It is clear that this occurs along the equilibrium path, but why is an agent in a $\bar{\beta}$ job unable to profit by choosing $(\underline{a}, \bar{\theta})$? This would appear to yield an immediate extra payoff of $\bar{a} - \underline{a}$ (at a cost of $\bar{\theta}$) and also to convince the principal that the job is actually $\underline{\beta}$, allowing the agent to choose \underline{a} in the second period and produce y_2 for an additional extra period-two payoff of $\bar{a} - \underline{a}$. The total extra payoff of $2(\bar{a} - \underline{a})$ exceeds the extra cost of $\bar{\theta}$ [cf. (4)], apparently making this optimal. However, job switching ensures that the agent will occupy a different job in the next period, with an equilibrium utility that depends upon information about the new job's productivity (which is revealed by the actions of the period-one agent in that job and is exogenous to this agent's calculation) and which is set equal to zero. There is then no period-two payoff from convincing the principal the job is actually a $\underline{\beta}$ job.

The expected profits in this case are then

$$\pi_1 = p_1[(y_1 - 2\bar{\theta} - \bar{a}) + (y_1 - \bar{a})] + (1 - p_1) \\ \times [(y_2 - \bar{a} - \bar{\theta}) + (y_2 - \bar{a})]$$

where the first bracketed expression gives profits if a job has high productivity (which is then weighted by the probability of such a job, or p_1), while the second bracketed expression gives profits if a job has low productivity (weighted by $1 - p_1$). Within each bracketed term, the parentheses indicate the net revenues for each period.

REFERENCES

- Aganbegyan, Abel, *The Economic Challenge of Perestroika*, Bloomington, IN: Indiana University Press, 1988.
- Amann, Ronald and Cooper, Julian, eds., *Industrial Innovation in the Soviet Union*, New Haven, CT: Yale University Press, 1982.
- _____ and _____, eds., *Technical Progress and Soviet Economic Development*, New York: Blackwell, 1986.
- Berliner, Joseph, *Factory and Manager in the USSR*, Cambridge, MA: Harvard University Press, 1957.
- _____, *The Innovation Decision in Soviet Industry*, Cambridge, MA: MIT Press, 1976.
- _____, "Organizational Restructuring of the Soviet Economy," in U.S. Congress, Joint Economic Committee, eds., *Gorbachev's Economic Plans*, Vol. 1, Washington, DC: U.S. Government Printing Office, 1987.
- Berry, M. J., "Towards an Understanding of R and D Innovation in a Planned Economy: The Experience of the Machine Tool Industry," in R. Amman and J. Cooper, eds., *Industrial Innovation in the Soviet Union*, New Haven, CT: Yale University Press, 1982, 39-100.
- Dearden, James A., Ickes, Barry W. and Samuelson Larry, "To Innovate or Not to Innovate: Incentives and Innovation in Hierarchies," working paper, Department of Economics, Pennsylvania State University, 1989.
- Dyker, David, *The Future of the Soviet Eco-*

- nomic Planning System*, New York: Sharpe, 1985.
- Ericson, R. E., "The Soviet Statistical Debate: Khanin vs. TsSU," Harriman Institute Occasional Paper 1, Columbia University, 1988.
- Freixas, X., Guesnerie, R. and Tirole, J., "Planning Under Incomplete Information and the Ratchet Effect," *Review of Economic Studies*, April 1985, 52, 173-92.
- Grossman, Gregory, "The 'Second Economy' of the USSR," in Morris Bornstein, ed., *The Soviet Economy: Continuity and Change*, Boulder, CO: Westview Press, 1981, 71-93.
- Hanson, Phillip, *Trade and Technology in Soviet-Western Relations*, New York: Columbia University Press, 1981.
- Hart, Oliver and Holmström, Bengt, "The Theory of Contracts," in Truman F. Bewley, ed., *Advances in Economic Theory*, Cambridge: Cambridge University Press, 1987, 71-156.
- Hewett, Ed, *Reforming the Soviet Economy*, Washington: Brookings Institute, 1988.
- Holmström, Bengt, "Moral Hazard and Observability," *Bell Journal of Economics*, Spring 1979, 10, 74-91.
- _____, "Design of Incentives and the New Soviet Incentive Model," *European Economic Review*, February 1982, 17, 127-48.
- Ickes, Barry W. and Samuelson, Larry, "Job Transfers and Incentives in Complex Organizations: Thwarting the Ratchet Effect," *Rand Journal of Economics*, Summer 1987, 18, 275-86.
- Keren, M., Miller, J. and Thornton, J., "The Ratchet: A Dynamic Managerial Incentive Model of the Soviet Enterprise," *Journal of Comparative Economics*, December 1983, 7, 347-67.
- Leary, Neil A. and Thornton, Judith, "Are Socialist Industries Insulated Against Innovation? A Case Study of Technological Change in Steelmaking," *Comparative Economic Studies*, January 1989, 31, 42-65.
- Ma, Ching-to, "Unique Implementation of Incentive Contracts with Many Agents," STICERD working paper 87/146, London School of Economics, 1987.
- _____, "Implementation in Dynamic Job Transfers," *Economics Letters*, 1988, 24(4), 391-95.
- _____, Moore, John and Turnbull, Stephen, "Stopping Agents from 'Cheating,'" *Journal of Economic Theory*, December 1988, 46, 355-72.
- Manevich, E., "Means of Restructuring the Economic Mechanism," *Problems of Economics*, May 1987, 30, 81-97.
- Mirrlees, J., "Notes on Welfare Economics, Information, and Uncertainty," in M. Balch, D. McFadden, and S. Wu, eds., *Essays on Economic Behavior under Uncertainty*, Amsterdam: North-Holland, 1974, 243-57.
- Nove, A., *The Soviet Economic System*, London: Allen and Unwin, 1977.
- Popper, Steven W., "The Diffusion of Numerically Controlled Machine Tools in Hungary," in Josef Brada and Istvan Dobozi, eds., *The Hungarian Economy in the 1980's*, Greenwich, CT: JAI Press, 1988.
- Riordan, Michael H., "Hierarchical Control and Investment Incentives in Procurement," mimeo, Stanford University, 1987.
- Rutland, Peter, "The Shchekino Method and the Struggle to Raise Labour Productivity in Soviet Industry," *Soviet Studies*, July 1984, 36, 345-65.
- Scherer, F. M., *Innovation and Growth*, Cambridge, MA: MIT Press, 1984.
- Selyunin, V. and Khanin, G., "Lukavaya tsifra" ("The Cunning Figure"), *Novy Mir*, February 1987, 2, 181-201.
- Shavell, Steven, "Risk Sharing and Incentives in the Principal and Agent Relationship," *Bell Journal of Economics*, Spring 1979, 10, 55-73.
- Tremblay, V. G., "Perestroika and Soviet Statistics," *Soviet Economy*, January 1988, 4, 65-94.
- Weitzman, M. L., "The 'Ratchet Principle' and Performance Incentives," *Bell Journal of Economics*, Spring 1980, 11, 302-8.

Intergenerational Income-Group Mobility and Differential Fertility

By C. Y. CYRUS CHU AND HUI-WEN KOO*

One question development economists are especially interested in, but so far left unanswered, is: how would the societal income distribution be affected by introducing a family-planning program to reduce the reproduction rate of the poor, which is usually high in developing countries? The purpose of this paper is to search for analytical answers to this question. We are able to make definite comparisons about some class of inequality measures of the steady-state societal income distributions, and these comparisons provide strong theoretical support in favor of the above-mentioned family-planning program. (JEL 112, 841)

In the past two decades, considerable attention has been given to the investigation of the relationship between population growth and the distribution of income. Many researchers have argued, based on their empirical findings, that population growth rate is positively related to income inequality.¹ However, Bryan Boulier (1982) and David Lam (1986b) have criticized these empirical results as too sensitive both to model specifications and to the selection of data sets. Moreover, because of the absence of a rigorous theoretical structure, there are difficulties both in interpreting the empirical evidence and in deducing persuasive policy implications from such evidence.

The difficulty associated with the theoretical modeling of the relationship between income distribution and population growth,

as Lam (1986a) pointed out quite correctly, hinges upon two factors: income-specific differential fertility and intergenerational income mobility. Indeed, when a distribution of income is referred to, it goes without saying that we are thinking about an economy with various income groups. As long as the reproduction rates or crude fertility rates across income groups are different, as is especially obvious in most developing countries, the property of income-specific differential fertility has to be embodied in the model. With differential fertility, population growth rate by definition becomes simply the weighted average of reproduction rates of all income groups, and therefore the compositional effects will confound the relationship between population growth and inequality. This suggests that the causal relationship between income distribution and population growth as a whole is not a very meaningful topic to analyze, and the key question that needs to be addressed instead is the relationship between income distribution and reproduction behavior of some particular income groups.

The second difficulty associated with the theoretical modeling is the concern of intergenerational income mobility. It is clear that the children of all income groups have the possibility of upward (or downward) mobility into other income groups. Since the societal distribution of income for any specific time period is formed by the income of its

*Chu: Department of Economics, National Taiwan University and Institute of Economics, Academia Sinica; Koo: Department of Economics, National Taiwan University, Taipei, Taiwan. We thank Brian Arthur at Stanford University for his continuous encouragement, which facilitated the completion of this paper. We are also indebted to Christophe Lefranc and two anonymous referees for their various comments and suggestions on an earlier version of this paper. Any remaining errors are our own responsibility.

¹See, for example, Irma Adelman and Cynthia Taft Morris (1973), James Kocher (1973), W. Rich (1973), Calman R. Winegarden (1978), Robert Repetto (1979), and Rati Ram (1984). For a detailed survey of the literature, see David Lam (1986b).

members, the stochastic evolution of parent-child income transition at the family level has inevitably rendered any discussion on the evolution of income distribution at the society level very complicated.

Recently, some progress has been made in modeling the complex interactions between income distribution and population growth. Lam (1986a) and Chu (1988, 1990) were able to characterize both differential fertility and intergenerational income mobility in a dynamic framework. The combination of the reproduction activity within each income-specific family and the dynamic income transition manifested in two generations of family members turns out to form a multitype Markov branching process, where the "type" refers to family income. Under certain regularity conditions, steady-state distribution of income and population growth rate have been shown to exist. Theoretically, the above model can then be used to evaluate the policy impact of changing the reproduction rate of one particular income group on the various properties of the steady-state income distribution.

Although the Markov branching process mentioned above has provided us with a neat structure for analyzing the dynamic interactions between income distribution and population growth, economists have so far been unable to derive any interesting policy implications from such a model. Recently, Lam (1986a p. 1110) has raised the following question: how would a different reproduction rate of the poor change various income-inequality measures in the steady state? Given the arguments of many economists that, in most developing countries, income inequalities are caused by high population growth² and that the high population growth rates in most developing areas are due to the high reproduction rate of the poor,³ his question clearly has policy rami-

fications. Unfortunately, even the simplest version of Lam's question—how would a different reproduction rate of the poor affect the proportion of poor in the steady state?—has not been answered satisfactorily.⁴ With even the simple question left unanswered, it is only logical that one hesitate before analyzing, or even posing, the more complicated but vital question: what impact will a different fertility rate of the poor have on the various inequality measures of the steady state?

The purpose of this paper is to search for analytical answers to the above question. The theoretical structure of this paper together with some elaborated discussion of background motivations are presented in Sections I and II. Section III provides the main theorem of this paper and its accompanying corollaries. The theorem essentially says that, as long as the income-specific reproduction rates in developing countries fit the stylized fact described in Ahluwalia (1976) and the intergenerational income-transition probability matrix satisfies the property of conditional stochastic monotonicity, which is a slight variant of the monotonicity conditions of G. I. Kalmykov (1962) and D. J. Daley (1968), then an increase in the reproduction rate of the poor will generate a steady-state income distribution that is conditionally first-degree stochastically dominated (CFSD) by the original distribution. The derived CFSD property implies the usual first-degree stochastic dominance (FSD) result of Josef Hadar and William R. Russell (1969), which enables us to compare some class of inequality measures of the societal income distribution in the steady state. Section IV shows that the elasticity of a change in the poor's fertility rate on the steady-state population growth can be easily calculated from the information of the income transition structure. Section V discusses various extensions and modifications of the results derived in Section III. Section VI provides a numerical example and demonstrates how the various comparative static results can be

²See Lam (1986a) for detailed references.

³As pointed out by M. S. Ahluwalia (1976 p. 326), "the most important link between population growth and income inequality is provided by the fact that different income groups grow at different rates, with the lower-income groups typically experiencing a faster rate of natural increase."

⁴See Chu (1987) for detailed explanation.

calculated, and the final section contains summaries and conclusions.

I. Theoretical Framework

In this section, we shall propose a theoretical structure to characterize the interaction between income-specific differential fertility and the intergenerational transmission of inequality. Although this interaction has been successfully characterized as a Markov branching process in Lam (1986a) and Chu (1987), very little discussion about the economic content behind the mathematical structure has been given. In what follows, we shall briefly present a household-decision framework in which parents' saving and fertility decisions are made endogenously. Furthermore, we will demonstrate how these family decisions are related to the Markov branching processes in question. It is believed that the present brief introduction will be helpful to the understanding of the insights of our later presentation.⁵

Let us consider the usual one-sex overlapping-generation framework with altruistic parents. Each individual lives two periods, young and old. At the beginning of people's old period, they receive bequests or other forms of endowments from their parents and from their own families. How a family head's income is related to the endowment he receives from his parent will be discussed below. We shall first discuss the decisions he is going to make *given* his family income y_t . The first decision a family head has to make is the number of children he wants to bear (F). After this is done, he has to divide his family income into family consumption (c_t) and savings $s_t = y_t - c_t$, which are further divided among children, either as human capital investment funds or as bequest.

Since each person may have different ability or luck, the relationship between a

child's earned income, denoted y_{t+1} , and his endowed capital (or bequests), denoted s_t/F , is not definite.⁶ Following Glenn Loury (1981), suppose we let a production function $\tilde{f}(\cdot, \cdot)$ summarize the interactions between a child's endowed capital, his luck or ability (denoted α_t), and his earned income (y_{t+1}):

$$(1) \quad y_{t+1} = \tilde{f}(s_t/F, \alpha_t).$$

In general, parents' optimal savings and optimal fertility are both dependent upon their family income, and hence we can rewrite (1) as

$$(2) \quad y_{t+1} = \tilde{f}(s(y_t)/F(y_t), \alpha_t) \\ \equiv f(y_t, \alpha_t)$$

where $s(\cdot)$ and $F(\cdot)$ respectively represent the optimal saving and fertility functions. Suppose α_t is independently identically distributed for all individuals in all periods and that $\partial f/\partial \alpha_t > 0$ (the marginal productivity of "luck" is always positive); then,

$$(3) \quad \Pr(y_{t+1} \leq y | y_t = x) \\ = \Pr(f(y_t, \alpha) \leq y | y_t = x) \\ = \Pr(\alpha \leq f^{-1}(x, y)) \\ = \mathcal{M}(y, x)$$

where f^{-1} is the inverse function of $f: f(x, f^{-1}(x, y)) \equiv y$. $\mathcal{M}(\cdot, \cdot)$ in (3) characterizes the cumulative transition probability function between y_t and y_{t+1} . Suppose the state space is discrete and there are n income classes in the economy with incomes $y^1 < y^2 < \dots < y^n$; then, we can relate the transition mass function $M_{ij} \equiv M(i, j)$ defined in Lam (1986a), which represents the probability that a child of income class j

⁵More details about the connection between the micro-level household decision and the macro-level branching process of the society can be found in Chu (1990).

⁶Here we do not consider the complication that parents may divide their bequests unevenly among children, either to compensate or to reinforce the ability differences of children (see e.g., Eytan Sheshinski and Yoram Weiss, 1982).

becomes a member of class i , with our cumulative transition function $\mathcal{M}(\cdot, \cdot)$:

$$\begin{aligned}\Pr(y_{t+1} \leq y^i | y_t = y^j) &\equiv \mathcal{M}(y^i, y^j) \\ &\equiv \sum_{i=1}^I M(i, j) \\ &\equiv \sum_{i=1}^I M_{ij}.\end{aligned}$$

What is demonstrated above is a micro foundation for the Markov branching process given in Lam (1986a). From (2) and (3), we see that there are two implicit elements in the specification of intergenerational transmission of income distribution: parents' optimal saving function $s(\cdot)$ and their optimal fertility function $F(\cdot)$. The former contains parents' implicit trade-offs between present family consumption and bequests or investments on children, and the latter characterizes a balance between parents' marginal benefit and marginal cost of childbearing. Factors that affect these two decisions will also affect the income transition structure and, hence, the steady-state income distribution. It should be noticed that parents' fertility decision and consumption/saving decision usually interact with each other. For example, if a parent is considering whether to have an additional child, he will note that this child will have a *capital-dilution* effect on per-child bequests [see (2)], and hence he may also want to consider a change in his saving decision. Thus, a policy that induces parents to reduce fertility sometimes will also cause a change in the income transition structure in (3).

In the following section, we show that the basic dynamic income transition structure characterized in (3) will eventually generate a steady state of income distribution, which is the main target of our later policy analysis. Various extensions and complications are discussed in Section V.

II. The Steady State of Income Distribution

Following the notation often adopted in the literature, denote the size of the n income groups in period t by $\mathbf{P}'_t =$

$[P_{1,t}, P_{2,t}, \dots, P_{n,t}]$, and the income-specific net reproduction rates by a diagonal matrix \mathbf{F} . Taking each period as a generation, Lam (1986a) showed that the evolution of income-specific population sizes would be characterized by the following identity:⁷

$$(4) \quad \mathbf{P}_t = \mathbf{M}\mathbf{F}\mathbf{P}_{t-1}$$

where the mobility matrix \mathbf{M} is defined in Section I. Let the proportion of the population in income group i at time t be $\pi_{i,t}$. Dividing both sides of (4) by $N_{t-1} \equiv \sum_{i=1}^n P_{i,t-1}$, the total population size at time $t-1$, yields

$$\begin{aligned}(5) \quad \mathbf{M}\mathbf{F}\boldsymbol{\pi}_{t-1} &= \mathbf{M}(\mathbf{P}_{t-1}/N_{t-1}) \\ &= \mathbf{P}_t/N_{t-1} \\ &= (\mathbf{P}_t/N_t)g_t = \boldsymbol{\pi}_t g_t\end{aligned}$$

or explicitly⁸

$$(5') \quad \sum_j F_j M_{ij} \pi_{j,t-1} = \pi_{i,t} g_t$$

where $g_t \equiv (N_t/N_{t-1})$ is the population growth rate at period t , and $\boldsymbol{\pi}'_t \equiv [\pi_{1,t}, \pi_{2,t}, \dots, \pi_{n,t}]$. Summing both sides of (5') over i and using the property $\sum_i M_{ij} = 1$ $\equiv \sum_i \pi_{i,t}$, we get

$$(6) \quad g_t = \sum_j F_j \pi_{j,t-1}.$$

In the steady state, (5) will become $\mathbf{M}\mathbf{F}\boldsymbol{\pi}^* = \boldsymbol{\pi}^* g^*$, where $\boldsymbol{\pi}^*$ and g^* denote the respective variables in the steady state.

Let \mathbf{T} be defined as $\mathbf{F}'\mathbf{M}'$; then T_{jj} represents the expected number of j -group chil-

⁷Equation (4) characterizes a Markov branching process. It should be noticed that the state variable of this branching process is not income but, rather, a point distribution. A typical point distribution at time t is written as $Z_t = (y_1, P_{1,t}; y_2, P_{2,t}; \dots; y_n, P_{n,t})$, which means that there are P_{it} people with income y_i , $i = 1, \dots, n$ at time t . As $t \rightarrow \infty$, the state variable Z_t will not converge by itself; but the *proportion* of people in various income classes will converge. See Theodore Harris (1963 Ch. III) for detailed explanation.

⁸To simplify our notations, the summation sign in this paper always ranges from 1 to n unless otherwise specified.

dren born to an i -group parent. On the assumption that the offspring of all income groups have positive probability of joining other income groups, and assuming that $F_i > 0$ is finite for all i , the matrix T satisfies the requirements of positive regularity. The following theorem proved by Charles Mode (1970 pp. 14, 30; see also Samuel Karlin and Howard Taylor, 1975 p. 546) is essential to our later discussion.⁹

THEOREM 1: *If T is positively regular, then T has a unique positive dominant eigenvalue g . Corresponding to g , there are right and left eigenvectors $\mu' = (\mu_1, \dots, \mu_n)$ and $\nu = (\nu_1, \dots, \nu_n)$, having strictly positive elements and with the properties $g\nu = \nu T$, $T\mu = g\mu$, and $\nu\mu = 1$. If $g > 1$, the frequency distribution $\pi_{i,t}$ given in (5') will converge to $\pi_i^* \equiv \nu_i / (\nu_1 + \dots + \nu_n) \forall i$ almost surely. Furthermore, if we set $T_1 = \mu\nu = (\mu_i\nu_j)$, it follows that $T^n/g^n \rightarrow T_1$.*

Mode's theorem tells us that as long as T is positively regular, starting with any arbitrary vector $\pi_0 \equiv [\pi_{1,0}, \dots, \pi_{n,0}]$, the iteration of equation (5) will generate the steady-state π^* . This result will be particularly useful for our analysis in the following sections. It is also shown by Mode (1970) that if $g \leq 1$, then the population will become extinct with probability one. Thus, the case $g > 1$ is the only interesting case to discuss here.¹⁰ In what follows we study how the steady-state distribution π^* will change as the poor's reproduction rate changes.

III. Distributional Impact of Changing Income-Specific Reproduction Rates

The question Lam (1986a) addressed but failed to answer was: what is the sign of $\partial\pi_1^*/\partial F_1$? With the sign of $\partial\pi_1^*/\partial F_1$ unknown, it does seem extremely difficult to analyze the impact of a changing F_1 on the various steady-state inequality measures, which are usually nonlinear functions of *all*

elements of the steady-state π^* . This difficulty has prompted us to search for other ways to tackle the problem. In a seminal paper, Anthony Atkinson (1970) pointed out that, instead of comparing various inequality measures with different and specific underlying concepts of social welfare, it is more appropriate to apply the notion of stochastic dominance and compare the distribution of income directly, with *minimum* restrictions on the properties of social welfare function. This is basically what we intend to do below.

As in Section II, let us, without loss of generality, order the indexes (subscripts) of all income groups in such a way that $y_1 \leq y_2 \leq \dots \leq y_n$. We shall put forth two assumptions.

ASSUMPTION 1:

$$F_1 \geq F_2 \geq \dots \geq F_n.$$

ASSUMPTION 2:

$$\frac{\sum_{i=1}^I M_{i1}}{\sum_{j=1}^J M_{j1}} \geq \frac{\sum_{i=1}^I M_{i2}}{\sum_{j=1}^J M_{j2}} \geq \dots \geq \frac{\sum_{i=1}^I M_{in}}{\sum_{j=1}^J M_{jn}}$$

$$1 \leq I \leq J \leq n.$$

The first assumption characterizes a stylized fact in developing countries referred to by Ahluwalia (1976 p. 326) that "the lower income groups typically experienc[e] a faster natural rate of increase." The second assumption requires that the mobility matrix obeys the property of conditional stochastic monotonicity (CSM), which is a variant of Kalmykov's (1962) condition of stochastic monotonicity (SM):

$$(SM) \quad \sum_{i=1}^I M_{i1} \geq \sum_{i=1}^I M_{i2} \geq \dots \geq \sum_{i=1}^I M_{in}$$

$$1 \leq I \leq n.$$

Notice that CSM implies CM but not vice versa, and hence CSM is a stronger assumption. In our context, Assumption 2 means that if a poor kid and a rich kid both fall into the poorest J classes, it is more likely

⁹The part of Mode's theorem that is irrelevant to our later presentation is not repeated here.

¹⁰See Chu (1988) for detailed discussion about the meaning of $g > 1$.

that the poor kid will be poorer than the rich kid, which seems to be an intuitively appealing statement. More discussion about the comparison between Assumption 2 and Kalmykov's SM condition will be left to Section V.

Now we shall introduce the main theorem of this paper. Suppose \mathbf{M} and \mathbf{F} are the original mobility and fertility matrices, and without loss of generality suppose at period zero the steady state associated with \mathbf{M} and \mathbf{F} has been reached with π_0 the steady-state income distribution. Let us consider a policy experiment that increases F_1 by δ . After such an increase in F_1 , π_0 obviously will no longer be the steady state, and a sequence of distribution vectors $\pi_1, \pi_2, \dots, \pi_t, \dots$ will evolve according to equation (5). The following theorem can be established:

THEOREM 2: *If Assumptions 1 and 2 hold, then*

$$(7) \quad \frac{\sum_{i=1}^I \pi_{it}}{\sum_{j=1}^J \pi_{jt}} \geq \frac{\sum_{i=1}^I \pi_{i0}}{\sum_{j=1}^J \pi_{j0}} \quad 1 \leq I \leq J \leq n \quad \forall t.$$

The proof is given in the Appendix.

When we set $J = n$ in the denominator of (7), we have

$$(8) \quad \sum_{i=1}^I \pi_{it} \geq \sum_{i=1}^I \pi_{i0} \quad 1 \leq I \leq n \quad \forall t$$

which is the conventional first-degree stochastic dominance (FSD) relation which was found by Jean-Pierre Danthine and John Donaldson (1981) while comparing distributions of two Markov (not branching) processes. Thus our *conditional* first-degree stochastic dominance (CFSD) result in (7) is stronger than the previous unconditional FSD result. In order to make comparisons with the existing literature, in what follows we will concentrate on the implications of the FSD result in (8).

When we let t go to infinity in (8), π_t on the left-hand side will converge to a new steady-state distribution (denoted π_*), as

asserted in Theorem 1. Thus, we have

$$(9) \quad \sum_{i=1}^I \pi_{i*} \geq \sum_{i=1}^I \pi_{i0} \quad 1 \leq I \leq n$$

from which many interesting economic implications can be derived. First, by setting $I = 1$ in (9), we obtain

$$\pi_{1*} \geq \pi_{10}.$$

When δ (the introduced fertility difference between regimes $*$ and 0) is infinitesimal, we have the usual comparative static result:

$$\frac{\partial \pi_1}{\partial \delta} = \lim_{\delta \rightarrow 0} \frac{\pi_{1*} - \pi_{10}}{(F_1 + \delta) - F_1} > 0.$$

This provides an answer to the question raised by Lam (1986a p. 1109): under what conditions can we determine the sign of $\partial \pi_1 / \partial F_1$? From the analysis of Chu (1987), it is clear that the conditions that enable us to answer Lam's question must include information on all the terms of matrices \mathbf{M} and \mathbf{F} . The question then becomes one of whether the conditions we propose are reasonable and whether they contain clear economic interpretations. From the discussions presented at the beginning of this section, we believe that Assumptions 1 and 2 satisfy these two criteria.

Another similar property that can be derived from (9) concerns the comparative statics of the proportion rich in the steady state. Since $\sum_{j=1}^n \pi_{j*} = 1 = \sum_{j=1}^n \pi_{j0}$, by setting $I = n - 1$ in (9) we have

$$\pi_{n*} \leq \pi_{n0}.$$

Thus, reducing the poor's reproduction rate will definitely increase the steady-state proportion rich.

Furthermore, the FSD result in (9) also allows us to analyze the welfare impact of a changing F_1 . For the class of Benthamite social welfare functions

$$W_k = \sum_j U(y_j) \pi_{jk}$$

with monotonically increasing $U(\cdot)$, it has long been understood (see Hadar and Russell, 1969 theorem 1) that π_0 exhibits FSD to π_* implies

$$(10) \quad \sum_j U(y_j) \pi_{j0} = W_0 \geq W_* \\ = \sum_j U(y_j) \pi_{j*}.$$

Therefore, one can conclude that, if Assumptions 1 and 2 are satisfied, then a reduction in F_1 can increase social welfare for a very large class of social welfare functions. As pointed out by Atkinson (1970), the above-mentioned direct ranking of income distribution, which does not rely on detailed functional specifications of social welfare, is a better alternative to conventional comparisons of various (perhaps conflicting) inequality measures.

Finally, if we set $U(X) = X \forall X$ in the definition W_k , then one particular implication of (10) will be that the mean of the steady-state income distribution will also be increased by a reduction in F_1 .

In summary, the above discussion shows that the high reproduction rate of the low-income group has a negative effect on all the distributional measures that can be explicitly worked out. These findings are strong theoretical support for family-planning programs that advocate a lower reproduction rate of the poor in developing countries.

IV. The Impact of Changing Income-Specific Reproduction Rate on Steady-State Population Growth

Having analyzed the distributional impact of a changing F_1 , we now study how the steady-state population growth rate will be affected by changes in F_1 . For demonstration purpose, let us designate the reproduction rate of the poor as $F_1 = F_1^* + \delta$ and let δ characterize the change in F_1 . For any given δ , the dynamic transition rule in (5') can be rewritten as

$$(11) \quad g_t(\delta) \pi_{i,t}(\delta) = \sum_j T_{ji}(\delta) \pi_{j,t-1}(\delta)$$

where $T_{ji} \equiv F_j M_{ij}$, as demonstrated in Section II, and where we let the δ 's that follow

all variables remind us that the dynamic system (11) is affected by the variable δ . Differentiating (11) with respect to δ and evaluating the result at $\delta = 0$ yields

$$(12) \quad g_t(0) \frac{d\pi_{i,t}(0)}{d\delta} \\ = -\pi_{i,t}(0) \frac{dg_t(0)}{d\delta} \\ + \frac{T_{1i}(0)}{F_1^*} \pi_{1,t-1}(0) \\ + \sum_j T_{ji}(0) \frac{d\pi_{j,t-1}(0)}{d\delta}.$$

Equation (10) characterizes the comparative dynamics of an infinitesimal change of F_1 at the point F_1^* . Suppose that at period zero $\delta = 0$ and that the steady state corresponding to $\delta = 0$ (or $F_1 = F_1^*$) has been achieved; then, by the definition of the steady state, we have $\pi_{i,t}(0) = \pi_i^* \forall t$ and $g_t(0) = g^* \forall t$. Thus, equation (12) can be written as

$$(13) \quad g^* \frac{d\pi_{i,t}}{dF_1} = -\pi_i^* \frac{dg_t}{dF_1} + \frac{T_{1i}}{F_1^*} \pi_1^* \\ + \sum_j T_{ji} \frac{d\pi_{j,t-1}}{dF_1}$$

where we drop the "(0)" in each term to simplify our later presentation. One can iteratively lag (13) one period and substitute the lagged result in the last term on the right-hand side of (13) to obtain¹¹

$$(14) \quad g^* \left(\frac{d\pi_{i,t+1}}{dF_1} - \frac{d\pi_{i,t}}{dF_1} \right) \\ = -\pi_i^* \frac{dg_{t+1}}{dF_1} + \frac{\pi_1^*}{F_1^*} \cdot \frac{T_{1i}^{t+2}}{(g^*)^{t+1}} \\ + \sum_j \frac{T_{ji}^{t+1} \cdot (d\pi_{j,0}/dF_1)}{(g^*)^t} \\ - \sum_j \frac{T_{ji}^t \cdot (d\pi_{j,0}/dF_1)}{(g^*)^{t-1}}$$

¹¹Technical detail is available from the authors upon request.

where $T_{ji}^{t+1} = \sum_k T_{jk}^t T_{ki}$. The almost-sure convergence property of the dynamic system characterized in Theorem 1 tells us that g and π will eventually converge to a new steady state, and therefore the changes in g and π will also converge to constants: $dg_t/dF_1 \rightarrow dg^*/dF_1$ and $d\pi_{i,t}/dF_1 \rightarrow d\pi_i^*/dF_1$. Furthermore, Theorem 1 tells us that as $t \rightarrow \infty$, $T_{ji}^t \rightarrow (g^*)^t \mu_j \nu_j$, which implies that the last two terms of (14) cancel each other out. With the above information, (14) can be further simplified as

$$(15) \quad 0 = -\pi_i^* \frac{dg^*}{dF_1} + \frac{\pi_1^*}{F_1} \mu_1 \nu_i g^*.$$

Because $\pi_i^* = \nu_i / \sum_j \nu_j \forall i$ by Theorem 1, (15) can be rearranged to

$$(16) \quad \left(\frac{dg^*}{dF_1} \right) \left(\frac{F_1}{g^*} \right) = \mu_1 \nu_i > 0.$$

It is worth noting from the above derivation that the validity of (16) is independent of Assumptions 1 and 2 and the source of the income-specific fertility change; hence, what we actually prove is the following.

THEOREM 3:

$$(dg^*/dF_1) \cdot (F_i/g^*) = \mu_i \nu_i > 0 \\ i = 1, \dots, n.$$

Theorem 3 is an easy formula that enables us to predict the steady-state change of population growth as a result of changes in the reproduction rate of any income group. Furthermore, since $\sum_j \mu_j \nu_j = 1$ by Theorem 1, it is clear that $\mu_i \nu_i < 1$ for $i = 1, \dots, n$. Thus, Theorem 3 signifies that a reduction in F_i will entail a corresponding decrease in the steady-state population growth rate, but the elasticity of changing is less than one.

V. Discussion and Extensions

A. Comparative Dynamics

What is emphasized in Section III is that a decrease in F_1 will make the steady-state income distribution exhibit CFSD to the

original distribution. However, this in itself is not sufficient for us to argue for the soundness of the policy of reducing F_1 without knowing how income distributions would change in transition periods. The answer to the latter is provided in (7), in which we see that as F_1 decreases, π_i would always exhibit CFSD to π_0 in all transition periods until the steady state is reached. This completes the comparative dynamics of our analysis and further strengthens our confidence in family-planning programs for the poor.

B. Family Size and Intergenerational Mobility

In Lam (1986a) as well as in our discussions in Section III, it is assumed that a reduction in F will not give rise to any changes in the terms of the mobility matrix. However, many studies have also found a persistently negative relationship between family size and child achievements in general.¹² In particular, it has been argued that, with a high F_1 , a poor family would have less to spend in per capita human capital investment on children, which would negatively affect the upward mobility of these children. Thus, as F_1 changes, we also expect changes in the first column of \mathbf{M} , which embodies the information of the upward mobility of the poor.

Denote $(\mathbf{F}^A, \mathbf{M}^A)$ and $(\mathbf{F}^B, \mathbf{M}^B)$ respectively as the \mathbf{F} and \mathbf{M} matrices before and after the change of F_1 . The change from $(\mathbf{F}^A, \mathbf{M}^A)$ to $(\mathbf{F}^B, \mathbf{M}^B)$ can be explored in two different stages: (i) $(\mathbf{F}^A, \mathbf{M}^A)$ to $(\mathbf{F}^B, \mathbf{M}^A)$ and (ii) $(\mathbf{F}^B, \mathbf{M}^A)$ to $(\mathbf{F}^B, \mathbf{M}^B)$. The first stage, which involves a changing F but with the mobility matrix unchanged, has already been analyzed in Section III, and now it is the second-stage change that we examine. Given the definite effect of changes in F_1 on the first column of \mathbf{M} , denoted \mathbf{M}_1 , it is pertinent to ask how such effect is going to take place. Let $\mathbf{M}_1' = [M_{11}, M_{12}, \dots, M_{1n}]$. Suppose that an increase in F_1 will make \mathbf{M}_1 change to $\bar{\mathbf{M}}_1$. We assume that the $\bar{\mathbf{M}}_1$ vector satisfies the following.

¹²For a general survey of studies on this topic, see Elizabeth M. King (1986).

ASSUMPTION 3:

$$\frac{\sum_{i=1}^I M_{i1}}{\sum_{j=1}^J M_{j1}} \leq \frac{\sum_{i=1}^I \bar{M}_{i1}}{\sum_{j=1}^J \bar{M}_{j1}} \quad 1 \leq I \leq J \leq n.$$

This means that, as the reproduction rate of the poor increases, each child of the poor will have a higher conditional probability of becoming poorer. More explicitly, given that a poor kid will fall into the poorest J class both before and after the change in \mathbf{M}_1 (caused by an increase in F_1), Assumption 3 says that he is more likely to be poorer after F_1 increases. Assumption 3 is also intuitively appealing.

THEOREM 4: *If Assumptions 1 and 2 hold and if the increase of F_1 worsens the upward mobility of the poor in such a way that Assumption 3 is satisfied, then inequality (7) will hold.*

PROOF:

We have shown in the proof of Theorem 2 that, under Assumptions 1 and 2, a policy change from $(\mathbf{F}^B, \mathbf{M}^A)$ to $(\mathbf{F}^B, \mathbf{M}^B)$ will make the transmission of the CFSD property hold in all *transition* periods. Thus, all that is needed to establish Theorem 4 is to show that the change from $(\mathbf{F}^B, \mathbf{M}^A)$ to $(\mathbf{F}^B, \mathbf{M}^B)$ will provoke a CFSD change in income distribution in the *first* period (see Case 1 of Appendix), and this is easy to establish under Assumption 3.

C. Immigration and Emigration

In the discussion of Section III, we did not consider the possible effects of migration. Poor parents in rural areas may want to send their children to urban areas, where better job opportunity is expected. Rich parents in developing countries may want to send their children to developed countries to evade the unstable political and economic environment of their homeland. This kind of income-specific emigration has two different impacts on the dynamic income-transition structure of the outflowing country: First, if x percent of the i -class children were to migrate out when they grow up, the net reproduction rate of the i th class could

become $\bar{F}_i = F_i(1 - x)$. Second, sending out some children may improve the capital-dilution effect within a family and hence change the i th class's mobility vector from \mathbf{M}_i to, say $\bar{\mathbf{M}}_i$. Similar analysis can be applied to the new dynamic process generated by $\bar{\mathbf{F}}$ and $\bar{\mathbf{M}}$.

The case of immigration is more difficult to analyze because the behavior and pattern of immigrants are exogenous to our model. Suppose $\mathbf{x}' \equiv [x_1, \dots, x_n]$ is the vector of total population that migrate to a society; then, the equation of motion in (4) will become $\mathbf{P}_t = \mathbf{MFP}_{t-1} + \mathbf{x}_t$. Once the relationship between \mathbf{x}_t and \mathbf{P}_{t-1} is specified, this augmented dynamic structure can also be studied in a similar fashion.¹³

D. The Age Structure

The overlapping-generation structure proposed in Sections II and III condenses a person's life into two periods, which is clearly a simplification of human age structure. More complicated age structure can be embodied in our basic framework by applying the multidimensional life table. Let us define an augmented transition matrix $\bar{\mathbf{M}}$, where $\bar{M}_{ia,ja+1}$ describes the transition from income class i at age a to income class j at age $a + 1$. Similarly, we can also include both age and income as arguments in the fertility function. Lam (1986a) provided a more detailed discussion about this extension, and the ergodicity result of this augmented model was also shown to hold.

One advantage of bringing in the age structure is that the possible interactions between family members' age structure and family incomes will have a chance to appear in the model.¹⁴ The cost of including both

¹³Ergodic results in an age-specific branching process with immigration have been analyzed by Thomas Espenshade et al. (1982).

¹⁴For example, it is Chayanov's belief that there is a correlation between family size and farm size (and hence family income) and that this changes with the life cycle of the peasant family. Chayanov argues that this kind of life-cycle interactions is one of the important factors that affect the evolution of Russian peasant families. See, for example, David Grigg (1983) for more detailed discussion.

age and income as state variables is that the transition rules among various states will become very complex, and it does not seem likely that we can find plausible and simple restrictions of fertility and mobility functions upon which interesting comparative dynamics about the steady states can be derived. The overlapping-generation structure may be too highly simplified to guide our analysis, but it appears to be an appropriate benchmark for the issue we are interested in.

E. *Analysis of the Impact of Alternative Policies*

Although our analysis in Section III was confined to the impact of changes in F_1 , the same can also apply to changes in F_i for $i \neq 1$ and other policy experiments. From the Appendix, we see that as long as Assumptions 1 and 2 are true, the transmission of the CFSD property will hold from period one onward. Thus, for policy experiments applied to an old steady state at period zero, it is quite possible to establish an CFSD comparison of steady states as long as the policy experiment in question can make the first-period distribution of income (π_1) exhibit CFSD to the original steady state (π_0). Since all that is needed for such a comparison is the information at period zero, in the case of well-specified policy experiments, such information should not be hard to come by.

F. *Comparisons Between Assumption 2 and Kalmykov's SM Condition*

As Carl Futia (1982) pointed out, comparative statics about how the invariant distributions of a Markov process change when the transition probability function is altered are very difficult in general, and the only known results in related research require the SM assumption of Kalmykov (1962) and Daley (1968). Since the model discussed in this paper is a Markov branching process which is more general and more complex than Markov processes, it seems unreasonable to start analysis with any assumptions weaker than the SM condition. Notice that

the SM condition places stochastic dominance (SD) restrictions on columns of M , whereas our Assumption 2 places conditional stochastic dominance (CSD) restrictions on columns of M . We summarize the conclusions obtained so far in the first two rows of Table 1.

Two points should be mentioned here. First, we have been given an example in which both Assumption 1 and SM are satisfied but no comparative dynamic results can be obtained.¹⁵ This makes us believe that the SM condition, which works for Markov processes in Daley's study, is insufficient to derive similar results for Markov branching processes. However, it is easy to show that, if one is willing to replace Assumption 1 with a stronger version as in the third row of Table 1, then the comparative dynamic results can still be derived.¹⁶ Second, although Assumption 2 is intuitively appealing, as we argued in Section III, it indeed requires more restrictions on entries of the mobility matrix than the SM condition.¹⁷ It would be interesting to investigate whether any results can be obtained with assumptions different from the SD type, or whether a comparative static (instead of comparative dynamic) result can be obtained with the usual SM condition. These seem to be promising directions for future research.

VI. A Numerical Example

As a last part of our illustration, a numerical example is used to work out various comparative statics. In the process, we also introduce a method that can efficiently generate various steady-state results. Let us set

$$M = \begin{bmatrix} 0.45 & 0.25 & 0.10 & 0.05 & 0.00 \\ 0.25 & 0.40 & 0.20 & 0.15 & 0.10 \\ 0.15 & 0.20 & 0.35 & 0.25 & 0.20 \\ 0.10 & 0.10 & 0.25 & 0.35 & 0.30 \\ 0.05 & 0.05 & 0.10 & 0.20 & 0.40 \end{bmatrix}$$

¹⁵The example was provided by Christophe Lefranc.

¹⁶Details are provided in Chu (1989), which is available upon request.

¹⁷For instance, it is easy to verify that the example in Lam (1986a p. 1112) satisfies SM but not Assumption 2.

TABLE 1—ASSUMPTIONS AND CONCLUSIONS DERIVED

Source	Assumptions	Results
Kalmykov (1962)	1) $F_1 = F_2 = \dots = F_n$ 2) SD on columns of \mathbf{M}	SD relation established for comparative dynamics
Theorem 2 of this Paper	1) $F_1 \geq F_2 \geq \dots \geq F_n$ 2) CSD on columns of \mathbf{M}	CSD relation established for comparative dynamics
Chu (1989)	1) $F_1 \geq F_2 = F_3 = \dots = F_n$ 2) SD on columns of \mathbf{M}	SD relation established for comparative dynamics

TABLE 2—STEADY STATES OF THE NUMERICAL EXAMPLE

	Case A ($F_1 = 1.00, i = 1, \dots, 5$)	Case B ($F_1 = 1.01, F_i = 1.00, i = 2, \dots, 5$)
μ'	[0.447, 0.447, 0.447, 0.447, 0.477]	[0.453, 0.447, 0.446, 0.445, 0.445]
$\pi^{*'} $	[0.167, 0.228, 0.238, 0.220, 0.146]	[0.168, 0.229, 0.238, 0.220, 0.146]
g^*	1.000	1.002
$K \equiv \pi^{*'} \cdot \mu$	0.447	0.447
$\nu \equiv \pi^{*'} / K$	[0.374, 0.511, 0.533, 0.492, 0.327]	[0.376, 0.511, 0.533, 0.491, 0.326]
$\mu_1 \cdot \nu$	0.167	0.170
$\Delta g F_1 / (\Delta F_1 g^*)$	0.168	0.169

and the steady states will be calculated for two cases: (A) $F_1 = 1, i = 1, \dots, 5$ and (B) $F_1 = 1.01$ and $F_i = 1, i = 2, \dots, 4$; that is, case (B) expands the poor's reproduction rate by one percent. Given the above information, one can iterate (5') and (6) to generate the steady state (g^{A*}, π^{A*}) and (g^{B*}, π^{B*}) for each case. Furthermore, for any column vector not orthogonal to π^* , a pair of sequences is defined as follows:

$$(17a) \quad \mathbf{X} = \mathbf{M}\mathbf{y}_{t-1} \quad t \geq 1$$

$$(17b) \quad \mathbf{y}_t = (\mathbf{X}_t' \mathbf{X}_t)^{-1/2} \mathbf{X}_t \quad t \geq 0.$$

It has been shown that the sequence \mathbf{y}_t will converge to $\pm \mu$,¹⁸ where μ is the right eigenvector corresponding to g^* . Finally, we calculate $\pi^{*'} \mu \equiv K$ and set $\nu \equiv \pi^{*'} / K$. Since ν so obtained will satisfy the normalization condition $\nu \mu = 1$, the value $\nu_1 \mu_1$ can be used to calculate the elasticity presented in Theorem 3. The results are summarized in Table 2.

From (17b) and the property that $\mathbf{y}_t \rightarrow \pm \mu$, it is clear that the μ vector has been normalized with $\sum_i \mu_i^2 = 1$, which together with the condition $\nu \mu = 1$ in Theorem 1 uniquely determines the steady-state (μ, ν) pair. From Table 2 one can verify two things: (i) π^{A*} exhibits CFSD to π^{B*} , and (ii) $\mu_1^A \nu_1^A < (\Delta g / \Delta F_1) \cdot (F_1^A / g^{A*}) < (\Delta g / \Delta F_1) (F_1^B / g^{B*}) < \mu_1^B \nu_1^B$. It can be checked that, when the experimental reduction in F_1 gets smaller, the formula in Theorem 3 will provide a more accurate prediction of the change in population growth.

VII. Conclusions

This paper examines the comparative dynamic relationship between income distribution and the reproductive rate of the low-income group. The problem is an important one, because many economists have argued that income inequalities in developing countries are caused by high population growth and that high population growth rates in most developing areas are due to the high reproductive rate of the poor. Our main finding is that, under some fairly reasonable assumptions, a reduction in the re-

¹⁸See Mode (1970 p. 104) for detail.

productive rate of the poor will cause a conditional stochastic dominance improvement in income distribution in the steady state as well as in all the transition periods. This result provides us with very strong theoretical support in favor of family-planning programs that encourage the poor in developing countries to reduce their reproductive rate. Furthermore, we derive an easy formula for calculating the elasticity of a change in the poor's fertility rate on the steady-state population growth rate. Technically, the analysis in this paper is an extension of Kalmykov's work on Markov processes to multitype Markov branching processes.

APPENDIX

The following lemma is needed to prove (7).

LEMMA:

$$(A1) \quad \frac{pA + (1-p)E}{pB + (1-p)F} \geq \frac{qC + (1-q)E}{qD + (1-q)F}$$

if $\frac{A}{B} \geq \frac{C}{D} \geq \frac{E}{F}$ and $A \geq C$,

where $A, B, \dots, F > 0$, and $1 \geq p \geq q \geq 0$.

The proof of the above lemma is straightforward algebra and hence will not be presented here. Technical details are available from the authors for interested readers.

Now we will proceed to prove Theorem 2.

PROOF OF THEOREM 2:

Case 1, when $t = 1$:

$$\begin{aligned} & \frac{\sum_{i=1}^I \pi_{i,1}}{\sum_{j=1}^J \pi_{j,1}} \\ &= \frac{\pi_{1,0}(F_1 + \delta) \sum_{i=1}^I M_{i1} + \dots + \pi_{n,0} F_n \sum_{i=1}^I M_{in}}{\pi_{1,0}(F_1 + \delta) \sum_{j=1}^J M_{j1} + \dots + \pi_{n,0} F_n \sum_{j=1}^J M_{jn}} \\ &\geq \frac{\pi_{1,0} F_1 \sum_{i=1}^I M_{i1} + \dots + \pi_{n,0} F_n \sum_{i=1}^I M_{in}}{\pi_{1,0} F_1 \sum_{j=1}^J M_{j1} + \dots + \pi_{n,0} F_n \sum_{j=1}^J M_{jn}} \end{aligned}$$

by Assumption 2

$$= \frac{\sum_{i=1}^I \pi_{i,0}}{\sum_{j=1}^J \pi_{j,0}}.$$

Case 2, when $t > 1$: Assuming

$$(A2) \quad \frac{\sum_{i=1}^I \pi_{i,t-1}}{\sum_{j=1}^J \pi_{j,t-1}} \geq \frac{\sum_{i=1}^I \pi_{i,0}}{\sum_{j=1}^J \pi_{j,0}} \quad I \leq J$$

we want to show

$$\frac{\sum_{i=1}^I \pi_{i,t}}{\sum_{j=1}^J \pi_{j,t}} \geq \frac{\sum_{i=1}^I \pi_{i,0}}{\sum_{j=1}^J \pi_{j,0}} \quad I \leq J$$

or equivalently, to show

$$(A3) \quad \frac{\sum_{i=1}^I \pi_{i,t}}{\sum_{i=1}^{I+1} \pi_{i,t}} \geq \frac{\sum_{i=1}^I \pi_{i,0}}{\sum_{i=1}^{I+1} \pi_{i,0}} \quad I < n.$$

The left-hand side of (A3) is

$$(A4) \quad \frac{\pi_{1,t-1}(F_1 + \delta) \sum_{i=1}^I M_{i1} + \dots + \pi_{n,t-1} F_n \sum_{i=1}^I M_{in}}{\pi_{1,t-1}(F_1 + \delta) \sum_{i=1}^{I+1} M_{i1} + \dots + \pi_{n,t-1} F_n \sum_{i=1}^{I+1} M_{in}}$$

A constructive proof of (A3) is given below. First, consider only the first two terms in (A4):

$$(A5) \quad \frac{\pi_{1,t-1}(F_1 + \delta) \sum_{i=1}^I M_{i1} + \pi_{2,t-1} F_2 \sum_{i=1}^I M_{i2}}{\pi_{1,t-1}(F_1 + \delta) \sum_{i=1}^{I+1} M_{i1} + \pi_{2,t-1} F_2 \sum_{i=1}^{I+1} M_{i2}} = \frac{(\pi_{1,t-1} + \pi_{2,t-1}) F_2 A}{(\pi_{1,t-1} + \pi_{2,t-1}) F_2 B}$$

where

$$\begin{aligned} A &= \left(\frac{\pi_{1,t-1}}{\pi_{1,t-1} + \pi_{2,t-1}} \right) \left(\frac{F_1 + \delta}{F_2} \right) \sum_{i=1}^I M_{i1} \\ &\quad + \left(\frac{\pi_{2,t-1}}{\pi_{1,t-1} + \pi_{2,t-1}} \right) \sum_{i=1}^I M_{i2} \\ B &= \left(\frac{\pi_{1,t-1}}{\pi_{1,t-1} + \pi_{2,t-1}} \right) \left(\frac{F_1 + \delta}{F_2} \right) \sum_{i=1}^{I+1} M_{i1} \\ &\quad + \left(\frac{\pi_{2,t-1}}{\pi_{1,t-1} + \pi_{2,t-1}} \right) \sum_{i=1}^{I+1} M_{i2}. \end{aligned}$$

Similarly, the first two terms of the right-hand side of (A3) could be written as

$$(A6) \quad \frac{(\pi_{1,0} + \pi_{2,0})F_2C}{(\pi_{1,0} + \pi_{2,0})F_2D}$$

where

$$\begin{aligned} C &= \left(\frac{\pi_{1,0}}{\pi_{1,0} + \pi_{2,0}} \right) \left(\frac{F_1}{F_2} \right) \sum_{i=1}^I M_{i1} \\ &\quad + \left(\frac{\pi_{2,0}}{\pi_{1,0} + \pi_{2,0}} \right) \sum_{i=1}^I M_{i2} \\ D &= \left(\frac{\pi_{1,0}}{\pi_{1,0} + \pi_{2,0}} \right) \left(\frac{F_1}{F_2} \right) \sum_{i=1}^{I+1} M_{i1} \\ &\quad + \left(\frac{\pi_{2,0}}{\pi_{1,0} + \pi_{2,0}} \right) \sum_{i=1}^{I+1} M_{i2}. \end{aligned}$$

From Assumptions 1 and 2, the Lemma, and (A2), it is clear that

$$(A7a) \quad \frac{A}{B} \geq \frac{C}{D}$$

which gives (A5) > (A6), and

$$(A7b) \quad \frac{C}{D} \geq \frac{\sum_{i=1}^I M_{i3}}{\sum_{i=1}^{I+1} M_{i3}} \quad \text{and} \\ A \geq C \geq \sum_{i=1}^I M_{i3}.$$

Now, we will proceed to the case with the first three terms in (A3). The first three terms in the left-hand side of (A3) could be rewritten as

$$(A8) \quad \frac{(\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1})F_3A'}{(\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1})F_3B'}$$

where

$$\begin{aligned} A' &= \left(\frac{\pi_{1,t-1} + \pi_{2,t-1}}{\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1}} \right) \left(\frac{F_2}{F_3} \right) A \\ &\quad + \left(\frac{\pi_{3,t-1}}{\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1}} \right) \sum_{i=1}^I M_{i3} \\ B' &= \left(\frac{\pi_{1,t-1} + \pi_{2,t-1}}{\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1}} \right) \left(\frac{F_2}{F_3} \right) B \\ &\quad + \left(\frac{\pi_{3,t-1}}{\pi_{1,t-1} + \pi_{2,t-1} + \pi_{3,t-1}} \right) \sum_{i=1}^{I+1} M_{i3}. \end{aligned}$$

Similarly, the first three terms in the right-hand side of (A3) could be rewritten as

$$(A9) \quad \frac{(\pi_{1,0} + \pi_{2,0} + \pi_{3,0})F_3C'}{(\pi_{1,0} + \pi_{2,0} + \pi_{3,0})F_2D'}$$

where

$$\begin{aligned} C' &= \left(\frac{\pi_{1,0} + \pi_{2,0}}{\pi_{1,0} + \pi_{2,0} + \pi_{3,0}} \right) \left(\frac{F_2}{F_3} \right) C \\ &\quad + \left(\frac{\pi_{3,0}}{\pi_{1,0} + \pi_{2,0} + \pi_{3,0}} \right) \sum_{i=1}^I M_{i3} \\ D' &= \left(\frac{\pi_{1,0} + \pi_{2,0}}{\pi_{1,0} + \pi_{2,0} + \pi_{3,0}} \right) \left(\frac{F_2}{F_3} \right) D \\ &\quad + \left(\frac{\pi_{3,0}}{\pi_{1,0} + \pi_{2,0} + \pi_{3,0}} \right) \sum_{i=1}^{I+1} M_{i3}. \end{aligned}$$

Again, from (A2), Assumptions 1 and 2, and (A7), we have

$$(A10a) \quad \frac{A'}{B'} \geq \frac{C'}{D'}$$

which gives (A8) > (A9), and

$$(A10b) \quad \frac{C'}{D'} \geq \frac{\sum_{i=1}^I M_{i4}}{\sum_{i=1}^{I+1} M_{i4}} \quad \text{and} \\ A' \geq C' \geq \sum_{i=1}^I M_{i4}.$$

Expressions (A10) will facilitate the next step with four terms.

REFERENCES

- Adelman, Irma and Morris, Cynthia Taft, *Economic Growth and Social Equity in Developing Countries*, Stanford: Stanford University Press, 1973.
- Ahluwalia, M. S., "Inequality, Poverty and Development," *Journal of Development Economics*, December 1976, 3, 307-42.
- Atkinson, Anthony B., "On the Measurement of Inequality," *Journal of Economic Theory*, September 1970, 2, 244-63.
- Boulier, Bryan, "Income Distribution and Fertility Decline: A Skeptical View," *Population and Development Review*, December 1982, 8, 159-78.
- Chu, C. Y. Cyrus, "The Dynamics of Population Growth, Differential Fertility, and Inequality: Note," *American Economic Review*, December 1987, 77, 1054-6.
- _____, "An Income-Specific Stable Population Model: Theory and Potential Applications," in T. Paul Schultz, ed., *Research in Population Economics*, Vol. 6, London: JAI Press, 1988.
- _____, "Differential Fertility and Income Distribution: Comparative Dynamics of Multitype Branching Process," Institute of Economics, Academia Sinica Discussion Paper 8909, June 1989.
- _____, "An Existence Theorem on the Stationary State of Income Distribution and Population Growth," *International Economic Review*, February 1990, 31, 171-85.
- Daley, D. J., "Stochastically Monotone Markov Chains," *Z. Wahrscheinlichkeitstheorie verw. Geb.* 1968, 10, 305-17.
- Danthine, Jean-Pierre and Donaldson, John B., "Stochastic Properties of Fast vs. Slow Growing Economies," *Econometrica*, July 1981, 49, 1007-33.
- Espenshade, Thomas J., Bouvier, Leon F. and Arthur, W. Brian, "Immigration and the Stable Population Model," *Demography*, February 1982, 19, 125-39.
- Futia, Carl, "Invariant Distribution and The Limiting Behavior of Markov Economic Models," *Econometrica*, March 1982, 50, 337-408.
- Grigg, David, *The Dynamics of Agricultural Change*, New York: St. Martin's Press, 1983.
- Hadar, Josef and Russell, William R., "Rules for Ordering Uncertain Prospects," *American Economic Review*, March 1969, 59, 25-34.
- Harris, Theodore E., *The Theory of Branching Processes*, Berlin: Springer-Verlag, 1963.
- Kalmykov, G. I., "On the Partial Ordering of One-Dimensional Markov Processes," *Theory of Probability and Its Applications*, 1962, 7, 456-59.
- Karlin, Samuel and Taylor, Howard M., *A First Course in Stochastic Processes*, 2nd ed., New York: Academic Press, 1975.
- King, Elizabeth M., "The Effect of Family Size on Family Welfare: What Do We Know?" in D. G. Johnson and R. D. Lee, eds., *Population Growth and Economic Development*, Madison: University of Wisconsin Press, 1986.
- Kocher, James, *Rural Development, Income Distribution, and Fertility Decline*, New York: Population Council, 1973.
- Lam, David, (1986a), "The Dynamics of Population Growth, Differential Fertility, and Inequality," *American Economic Review*, December 1986, 76, 1103-16.
- _____, (1986b), "Distribution Issue in the Relationship Between Population Growth and Economic Development," in D. G. Johnson and R. D. Lee, eds., *Population Growth and Economic Development*, Madison: University of Wisconsin Press, 1986.
- Loury, Glenn C., "Intergenerational Transfers and the Distribution of Earnings," *Econometrica*, July 1981, 49, 843-67.
- Mode, Charles J., *Multitype Branching Processes: Theory and Application*. New York: Elsevier, 1971.
- Ram, Rati, "Population Increase, Economic Growth, Educational Inequality, and Income Distribution," *Journal of Development Economics*, April 1984, 14, 419-28.
- Repetto, Robert, *Economic Equality and Fertility in Developing Countries*, Baltimore: Johns Hopkins University Press, 1979.
- Rich, W., *Smaller Families Through Social and Economic Progress*, Washington: Overseas Development Council, 1973.
- Sheshinski, Eytan and Weiss, Yoram, "Inequality Within and Between Families," *Journal of Political Economy*, February 1982, 90, 105-27.
- Winegarden, C. R., "A Simultaneous-Equations Model of Population Growth and Income Distribution," *Applied Economics*, December 1978, 10, 319-30.

Productivity, Health, and Inequality in the Intrahousehold Distribution of Food in Low-Income Countries

By MARK M. PITT, MARK R. ROSENZWEIG, AND MD. NAZMUL HASSAN*

A model is formulated incorporating linkages among nutrition, labor-market productivity, health heterogeneity, and the intrahousehold distribution of food and work activities in a subsistence economy. Empirical results, based on a sample of households from Bangladesh, indicate that, despite considerable intrahousehold disparities in calorie consumption, households are averse to inequality. Furthermore, consistent with the model, the results also indicate that both the higher level and greater variance in the calories consumed by men relative to women reflect in part the greater participation by men in activities in which productivity is sensitive to health status. (JEL 824, 122, 850).

A prominent if not distinguishing feature of low-income countries that has been incorporated into many models of behavior in such settings is the proximity of average income levels to subsistence. Models of savings behavior (Mark Gersovitz, 1983) and wage determination (Harvey Leibenstein, 1957; Joseph Stiglitz, 1976; Partha Dasgupta and Debraj Ray, 1984), for example, have demonstrated the possibility that behavior at low income levels may be quite distinct from that observed when income levels are well above those required for survival. Low-income societies are also characterized by an occupational distribution in which activities requiring high levels of energy expenditure predominate, and a number of recent studies have shown that health and food consumption directly affect productivity and wage rates in low-income environments (John Strauss, 1986; Anil Deolalikar, 1988; Jere Behrman and Deolalikar, 1989). In a subsistence regime, the allocation of

food is thus particularly important, and the measurement of the overall level of inequality in low-income countries must take into account how households in such environments distribute food among their individual members.

One salient aspect of the distribution of food in low-income settings that has caught the attention of many social scientists is the disparity in nutrients received by women compared to men, particularly in South and West Asian societies.¹ One hypothesis that has been advanced is that gender-based nutrient inequality reflects disparities in labor-market opportunities between men and women in these settings, with the pecuniary returns to a household from the allocation of food to women being less than those for men. Indeed, some empirical studies have shown the existence of a relationship between sex differences in infant mortality rates and differences in labor-market participation rates between men and women (Pranab Bardhan, 1974; Rosenzweig and T. Paul Schultz, 1982). However, there is little direct evidence of a relationship between the actual intrahousehold distribution of food across individuals and labor-market activities; nor is there a clear theoretical linkage established between labor-market characteristics and patterns of intrahousehold

*Department of Economics, Brown University, Providence, RI 02912; Department of Economics, University of Pennsylvania, Philadelphia, PA 19104; and Institute of Food Sciences and Nutrition, University of Dhaka, Dhaka, Bangladesh, respectively. This research was funded in part by NIH grant HD21096. An earlier version of this paper was presented at workshops at Baruch College, Brown University, Pennsylvania State University, the University of Minnesota, Gadjah Mada University, the University of Rochester, and Yale University. We are grateful to the referees for helpful comments.

¹For an extensive review of the literature concerned with gender inequality and the intrahousehold distribution of food, see Behrman (1990).

TABLE 1—HOUSEHOLD DISTRIBUTIONS OF CALORIES BY AGE AND SEX IN BANGLADESH

Statistic	Age < 6			6 ≤ Age < 12			Age ≥ 12		
	Males	Females	$X^2(d.f.)$	Males	Females	$X^2(d.f.)$	Males	Females	$X^2(d.f.)$
Mean household calorie consumption	891	751	2.35 (217)	1,549	1,536	0.25 (220)	2,672	2,063	609.1 (465)
Mean household coefficient of variation	43.6	41.1	0.26 (38)	11.1	10.5	0.23 (29)	11.5	7.05	4.48 (143)

Source: Nutrition Survey of Bangladesh, 1981–2.

food allocation that may arise in low-income environments.²

Although attention has mainly focused on gender inequality in food allocation, if the relationship between healthiness and productivity differs across occupations and activities, the distribution of activities across individuals within gender classes should also be related to the intrahousehold distribution of foods. Table 1 presents the means of average household calorie consumption and the intrahousehold coefficient of variation in calorie consumption by age and sex for a probability sample of 345 households from 15 villages in Bangladesh.³ These figures show that, while there is a large (30 percent) and statistically significant difference in the average number of calories allocated to men and women aged 12 and above, there is no difference between sexes in mean calories consumed by children ages 7 through 11. For children aged 6 and below, boys on average receive more calories than girls, however. Gender differences in aver-

age calorie consumption are thus highly age-dependent. Table 1 also shows, more interestingly, that mean within-household inequality in food consumption, measured by the coefficient of variation, is 64 percent higher among males aged 12 and over than among females of the same age. Among children less than 12 years of age, however, inequality in calorie consumption among boys and girls is similar.⁴

Table 2 displays the distribution of activities ranked by their energy requirements, within the same sex and age groups. These figures demonstrate that stratification by activities also varies by age and sex and in large part parallels what is observed in Table 1 for calorie consumption. The similarity in energy intensity and diversity of activities exhibited by girls and boys in the below-six and 6-to-12 age groups mirror the similarity in the mean and variability in calorie consumption among boys and girls in those age groups exhibited in Table 1. Furthermore, the large disparities in participation rates in high-energy-intensive activities between men and women aged 12 and over are consistent with the gender differences in the variability of calorie consumption depicted in Table 1 for that age group.

Tables 1 and 2 are suggestive of a direct linkage between the type of work activities

²One important study of the intrahousehold distribution of nutrients (Behrman, 1988) finds no apparent link between expected labor-market opportunities and sex disparities in nutrient consumption. However, this study only considers the allocation of foods among children less than 13 years of age, a large proportion of whom do not participate in the labor market and, perhaps more importantly, among whom there may be little differentiation with respect to work activities. For this group, the link between the labor market and food consumption can only be indirect and is in any case not explicitly modeled.

³Later in this paper, we describe the characteristics of this data set. Calorie consumption in Bangladesh is a good indicator of overall nutrient consumption, given the simplicity of the Bangladeshi diet, as discussed in Section III. The sex- and age-specific coefficients of variation are computed only for those households with two or more individuals in each group.

⁴Similar patterns characterize the Indian village data used by Behrman (1988). He shows that there are no sex differences in average nutrient allocations for children younger than 13, the subset of the population he studies. However, using the same data set, we find that for individuals aged 13 and above, mean calorie consumption is 12 percent higher for males. The variance in consumption among males is 15.6 percent higher than it is among females in that age group. Both of these differences are statistically significant.

TABLE 2—PERCENTAGE ACTIVITY DISTRIBUTION BY ENERGY REQUIREMENTS, AGE, AND SEX

Energy requirement	Age < 6		6 ≤ Age < 12		Age ≥ 12	
	Males	Females	Males	Females	Males	Females
Insignificant	98.7	99.3	70.5	69.1	26.8	20.6
Light	1.3	0.7	28.8	25.6	22.6	8.5
Moderate	0	0	0	4.5	2.82	68.2
Very high	0	0	0.7	0.8	31.9	1.2
Exceptionally high	0	0	0	0	15.9	1.5
Sample size (<i>N</i>)	133	129	140	133	433	473
χ^2 (<i>d.f.</i>)	0.28 (1)		6.87 (3)		625.4 (4)	
<i>P</i> ^a	0.600		0.076		0.0001	

^aSignificance level (probability).

and the intrahousehold distribution of food consumption, but they of course do not explain the diversity in activities within gender groups. In this paper, we examine the relationship between the household distribution of foods and labor-market activities in the context of a model incorporating (i) linkages among food consumption, health, and labor-market productivity and (ii) individual heterogeneity in inherent or "endowed" healthiness. The model takes as given differences in the opportunities for work activities by gender; conditional on the circumscribed activities of women, it yields implications for how the distribution of individual health endowments and nutrition-productivity linkages influence the distribution of food and energy expenditure (effort) across individual members of a household and provides a method for measuring gender-based discrimination by the household. Section I presents the model. Section II discusses the methodology used to compute individual endowments, and Section III reports estimates, based on a sample of households from 15 villages in Bangladesh, of the effects of food consumption and activities on weight-for-height, the effects of health endowments on calorie consumption by sex and age, and the effects of endowments on activity choice and income.

The empirical results appear to be consistent with the hypothesis that work activity distributions substantially influence the intrahousehold distribution of food. In particular, the greater participation by men in

energy-intensive activities in which health status may importantly influence productivity is in part responsible for both the higher level of calories consumed by adult men and the greater variance in calories consumed among men compared to women. We are able to infer from our estimates, however, that households are averse to inequality in health outcomes, with men bearing slightly more of the "cost" of equalization than women as a consequence of their participation in activities requiring high energy levels.

I. Theory

To analyze the relationships among the distribution of food, health, and labor-market activities, we set out a framework describing the allocation of food and the choice of labor-market "effort" across heterogeneous individuals residing in integrated household units, defined by common objective functions. For simplicity we assume that there is only one food or nutrient. The model can be readily extended to incorporate multiple foods and nutrients with no alteration in its basic implications. The health status h_i^k of an individual i among a class of individuals k is assumed to be influenced by food consumption c_i and by effort e_i expended in some work activity. In general, the effects of these variables on health may be nonmonotonic. However, we assume that in a subsistence economy food augments health, while effort decreases

health (stamina) such that

$$(1) \quad h_i^k = h^k(c_i, e_i, \mu_i)$$

$$\frac{\partial h_i^k}{\partial c_i} > 0 \quad \frac{\partial h_i^k}{\partial e_i} < 0$$

where μ_i is the endowed health of an individual, that component of health influenced by neither consumption nor effort.

Effort is rewarded in the labor market, with the returns to effort increasing with health status. The wage rate, w_i^k for an individual i in the class of individuals k is given by the following:⁵

$$(2) \quad w_i^k = w^k(e_i, h_i)$$

$$\frac{\partial w_i^k}{\partial e_i}, \frac{\partial w_i^k}{\partial h_i} > 0 \quad \frac{\partial^2 w_i^k}{\partial e_i \partial h_i} > 0.$$

Individuals are assigned to classes (age, sex) by the characteristics of the health and effort wage functions so that every member of each class has the same h and w functions; individuals are individually differentiated by their health endowments μ_i , which are known to all family members.

Expressions (1) and (2) capture the essential assumption of the nutrition-wage literature: that food consumption augments labor-market productivity, presumably via health status. However, while the nutrition-based efficiency wage literature assumes a purely technological relationship between effort and health (or food consumption), here both food consumption and labor effort are choice variables.⁶ Moreover, we allow the wage function to differ across classes of individuals, which may result from their

allocation to particular sets of activities. Thus, for example, in India, few women engage in plowing; in Bangladesh, no women are observed pulling rickshaws. The relationship between health and the returns to effort are likely to be quite different in those activities in which both women and men participate. Indeed, Behrman and Deolalikar (1989) and David Sahn and Harold Alderman (1988), based on data from India and Sri Lanka, respectively, found that health (measured as weight-for-height) and calorie consumption had significant positive effects on the wage rates of men but not women.

The allocations of food and work effort across individuals in a household unit are determined from the solution to the maximization problem

$$(3) \quad \max_{c_i^k, e_i^k} U(h_1^k, \dots, h_{n_k}^k, c_1^k, \dots, c_{n_k}^k, e_1^k, \dots, e_{n_k}^k) \quad k = 1, \dots, m$$

subject to

$$(4) \quad v + \sum_k \sum_i w_i^k - p \sum_k \sum_i c_i^k = 0$$

and functions (1) and (2), where v = nonearned income and p is the price of the food good. In the household welfare function (3), it is assumed that increases in both health status and food consumption augment utility, while increases in work effort lower utility.

The necessary first-order conditions for the allocation or assignment of food and work effort to individual i of class k are

$$(5) \quad \left(\frac{\partial U}{\partial h_i^k} \right) \left(\frac{\partial h^k}{\partial c_i} \right) + \frac{\partial U}{\partial c_i^k}$$

$$= \lambda \left[p - \left(\frac{\partial w^k}{\partial h_i^k} \right) \left(\frac{\partial h^k}{\partial c_i} \right) \right]$$

$$(6) \quad \left(\frac{\partial U}{\partial h_i^k} \right) \left(\frac{\partial h^k}{\partial e_i} \right) + \frac{\partial U}{\partial e_i^k}$$

$$= -\lambda \left[\frac{\partial w^k}{\partial e_i} + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial h^k}{\partial e_i} \right) \right]$$

⁵We also assume that the marginal product of health vanishes if there is no effort, so that the second derivative of health in the wage function is zero. If (2) is quadratic, for example, then $w = h(e + \gamma e^2)$.

⁶We assume that work time is fixed (and set to unity) as is conventionally assumed in the nutrient-wage literature. It is possible to include home production activities in total work time, with (2) being replaced by a goods-production function, with no alteration in the basic implications of the model. Our data set contains no information on the amount of time allocated to any activity.

where λ = marginal utility of income. Condition (5) states that the marginal cost of allocating an additional unit of food to person i is lower the greater the extent to which health augments work efficiency. Thus, if the members of class l participate in activities for which the market returns to health are greater compared to those activities in which members of class k participate (who are otherwise identical), then on average class- l individuals will receive higher allocations of food than will class- k individuals. Since we assume for simplicity that work (whether market or nonmarket) time is the same for all individuals (there are few idle women in low-income countries), it is not market work time (or even the average wage rate) that matters for food allocation (as in Rosenzweig and Schultz [1982]), but the type of activity engaged in, as defined by the wage-effort-health association.

Within a class, the distribution of food and work effort across individuals will depend on the distribution of endowments. To highlight the roles of both health in the labor market and household preferences in influencing these distributions, assume that the endowment is additive in (1). Thus, differences in endowments do not influence the health returns to food consumption. Consider first a model, nested in (3), in which household income is maximized. The maximand is the left-hand side of expression (4), and the necessary first-order conditions are given by (5) and (6) with the left-hand side of each expression replaced by zero. In the income-maximizing model, the relationships between the endowment of an individual i in class k and that person's allocation of food and work effort are given by

$$(7) \quad \frac{dc_i^k}{d\mu_i^k} = \left[\left(\frac{\partial^2 w^k}{\partial e_i \partial h_i} \right) \left(\frac{\partial h^k}{\partial c_i} \right) + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial e_i \partial c_i} \right) \right] \times \Phi^{-1} \frac{\partial^2 w^k}{\partial e_i \partial h_i} > 0$$

$$(8) \quad \frac{de_i^k}{d\mu_i^k} = - \left[\left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial c_i^2} \right) \right] \times \Phi^{-1} \frac{\partial^2 w^k}{\partial e_i \partial h_i} > 0$$

where

$$\Phi = \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial c_i \partial c_i} \right) \times \left[\frac{\partial^2 w^k}{\partial e_i \partial e_i} + \left(\frac{\partial^2 w}{\partial e_i \partial h_i} \right) \left(\frac{\partial h^k}{\partial e_i} \right)^2 + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial e_i \partial c_i} \right) \right] - \left[\left(\frac{\partial^2 w^k}{\partial e_i \partial h_i} \right) \left(\frac{\partial h^k}{\partial c_i} \right) + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial e_i \partial c_i} \right) \right] \times \left[\frac{\partial^2 w^k}{\partial e_i \partial h_i} + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial^2 h^k}{\partial e_i \partial c_i} \right) \right] > 0.$$

As indicated by equation (8), under the income-maximization regime those individuals with greater endowments of health supply more effort, because health augments the labor-market returns to effort ($\partial^2 w^k / \partial e_i \partial h_i > 0$). More-endowed individuals also receive more food because food increases health, which increases the returns to effort, and because effort depletes health status; increased food consumption both compensates for and enhances the return from increased effort. Thus, those individuals exerting greater effort or in effort-intensive activities will also be consuming more food. Moreover, those classes of individuals in activities for which the returns to work effort are more sensitive to health status will be characterized by greater differences in food consumption (and effort) compared to an otherwise identical class of individuals with the same distribution of endowments. This is because the magnitude of the (positive) endowment-food-con-

sumption slope in (7) depends positively on the degree to which health augments market returns to effort.

In the utility-maximization model (3), the relationships among own endowments, food consumption, and work effort are given by

$$\begin{aligned}
 (9) \quad \frac{dc_i^k}{d\mu_i^k} &= \left[p - \left(\frac{\partial w^k}{\partial h_i^k} \right) \left(\frac{\partial h^k}{\partial c_i} \right) \right] \left(\frac{\partial h^k}{\partial c_i} \right)^{-1} \\
 &\quad \times \left[- \left(\frac{\partial^2 h^k}{\partial c_i^2} \right) (S_{c_i c_i}) + \frac{dc_i^k}{dv} \right] \\
 &\quad - (S_{c_i e_i}) \left(\frac{\partial^2 w^k}{\partial e_i \partial h_i} \right) + \left(\frac{dc_i^k}{dv} \right) \left(\frac{\partial w^k}{\partial h_i} \right) \\
 (10) \quad \frac{de_i^k}{d\mu_i^k} &= \left[\frac{\partial w^k}{\partial e_i} + \left(\frac{\partial w^k}{\partial h_i} \right) \left(\frac{\partial h^k}{\partial e_i} \right) \right] \\
 &\quad \times \left(\frac{\partial h^k}{\partial e_i} \right)^{-1} \left[- \left(\frac{\partial^2 h^k}{\partial e_i^2} \right) (S_{e_i e_i}) + \frac{de_i^k}{dv} \right] \\
 &\quad + (S_{e_i e_i}) \left(\frac{\partial^2 w^k}{\partial e_i \partial h_i} \right) + \left(\frac{de_i^k}{dv} \right) \left(\frac{\partial w^k}{\partial h_i} \right)
 \end{aligned}$$

where dc_i^k/dv and de_i^k/dv are income effects on food and effort, $S_{c_i c_i}$ and $S_{e_i e_i}$ are the Hicks-Slutsky compensated own substitution effects (negative and positive, respectively), and $S_{e_i c_i}$ is the Hicks-Slutsky cross-compensated substitution effect, which is negative if effort (a "bad") and food consumption are substitutes. The first of the three right-hand-side terms in (9) and (10) arises from the welfare function in (3). This term indicates that the relationships among own endowments, food consumption, and effort depend on the relative magnitudes of substitution and income effects. If income effects are small, then in the absence of labor-market returns, higher-endowed individuals receive less food and provide more labor-market effort. Some of their higher health is thus taxed away via both the food and effort allocations; low-endowment individuals are "compensated" for their low

endowments by higher food and lower effort allocations.

The last two right-hand-side terms in (9) and (10) arise because of the health-effort interaction in the labor market. Both of these terms are positive in the food-allocation equation (9), given that food is a normal good. Thus, the association between own endowments and food consumption will be algebraically higher the more strongly health augments the returns to effort. If women are barred (or refrain) from participating in activities in which health status strongly affects productivity, then compensation (reinforcement) with respect to food is more (less) likely than among men.

We note that defining compensation with respect to the sign of the relationship between own endowments and an individual-specific input, such as foods, can be misleading when more than one allocated good affects health status and welfare. An alternative method of gauging compensation, and of more meaningfully assessing the differential treatment of different classes of individuals by the household, is to examine the net change in health status associated with a change in endowment. This is given by (11) in the additive endowment case:

$$\begin{aligned}
 (11) \quad \frac{dh_i^k}{d\mu_i^k} &= 1 + \left(\frac{\partial h_i^k}{\partial c_i^k} \right) \left(\frac{dc_i^k}{d\mu_i^k} \right) \\
 &\quad + \left(\frac{\partial h_i^k}{\partial e_i^k} \right) \left(\frac{de_i^k}{d\mu_i^k} \right).
 \end{aligned}$$

If the sum of the last two terms in (11) is negative (positive), then compensation (reinforcement) with respect to health occurs for group k ; reinforcement with respect to foods is thus not inconsistent with a household's aversion to inequality in health status. Expression (11) may differ across groups; intergroup differences in (11) thus are a measure of net discrimination across groups by the household with respect to health that incorporates both food and effort allocations.

While it is clear that the signs of the own endowment effects on food and effort do

not necessarily distinguish between the income- and welfare-maximizing models, the existence of cross endowment effects can only arise when household welfare is being maximized (and the welfare function is not linear in its arguments). It is straightforward to show that, although such effects cannot be signed in general, the cross effect of j 's endowment on i 's food consumption is more negative (given that the consumption of i and j are substitutes in the household welfare function) the stronger is the relationship between health and effort productivity for j .⁷ Thus, the cross effect of a woman's endowment on a man's food consumption, given the gender differences in activities exhibited in many South Asian societies, will be algebraically greater than will the effect of a change in a man's endowment on the woman's food allocation, while own endowment effects will be algebraically greater for males. Knowledge about both the health technology and the role of health in augmenting productivity is thus critical for understanding the determinants of the allocation of foods and effort levels.

II. Estimating the Relationships between Endowments and Household Resource Allocations

To estimate the association among the endowments of members of a household, their food consumption, and their expendi-

⁷The change in the endowment of individual j in class l on the food consumption of individual i in class k in the welfare-maximization model, is given by

$$\begin{aligned} \frac{dc_i^k}{d\mu_j^l} = & \left[p - \left(\frac{\partial w^l}{\partial h_j} \right) \left(\frac{\partial h^l}{\partial c_j} \right) \right] \left(\frac{\partial h^l}{\partial c_j} \right)^{-1} \\ & \times \left[- \left(\frac{\partial^2 h^l}{\partial c_j \partial c_j} \right) (S_{c_j c_j}) + \frac{dc_i^l}{dv} \right] \\ & + (S_{c_j e_i}) \left(\frac{\partial^2 w^l}{\partial e_j \partial h_j} \right). \end{aligned}$$

The cross effect of j 's endowment on i 's food consumption is more negative (given that e_j and c_j are substitutes in the household welfare function) the stronger is the relationship between health and effort productivity for j .

ture of effort, we employ a method first used in Rosenzweig and Schultz (1983), in which the health technology (1) is estimated directly, and based on the technology parameter estimates and the actual resources consumed or expended by each individual, individual-specific endowments are computed. There are two problems with this "residual" endowment method. First, if endowments, which are not directly observed by the researcher, influence resource allocations, consistent estimates of the household production technology cannot be obtained using least squares; that is, c_i , e_i , and the unobserved μ_i will be correlated in (1). One method of identifying the technology is to use instruments. In this case, food prices, labor-market variables reflecting labor demand, and exogenous components of income determine resource allocations but do not directly affect health status, given food and activity levels.

A second, less well-recognized problem with extracting estimates of endowments from estimates of the technology, which arises even when the technology is estimated consistently, is that the derived endowments will be measured with systematic error. Contrary to the assertions in Rosenzweig and Schultz (1983), the measurement errors in estimated endowments are not likely to be random, because the technology inputs, in this case individual-specific levels of nutrients, are unlikely to be measured without error. Endowment effects estimated by least squares are thus unlikely to be consistent, and the biases cannot necessarily be signed a priori.

To see the measurement-error problem and one solution, assume for simplicity that the health-production function contains only one nutrient (calories). The true (measured without error) endowment μ_i^* is thus

$$(12) \quad \mu_i^* = H_i^* - C_i^* \Gamma \quad i = 1, \dots, n$$

where H_i^* and C_i^* are the (unobserved) true values of health and calorie consumption, respectively, and Γ is the calorie effect on health. Assume that the observed values H and C have measurement errors u_i and e_i with classical errors-in-variables proper-

ties; that is,

$$(13) \quad H_i = H_i^* + u_i$$

$$(14) \quad C_i = C_i^* + e_i$$

where $E(H_i^*, u_i) = 0$, $E(C_i^*, e_i) = 0$, and $E(u_i, e_i) = 0$. Therefore, the estimated endowment $\hat{\mu}_i$ is

$$(15) \quad \hat{\mu}_i = (H_i^* + u_i) - (C_i^* + e_i)\hat{\Gamma} \\ = \mu_i^* + u_i - e_i\hat{\Gamma}$$

where $\hat{\Gamma}$ is the two-stage least-squares estimate of the calorie effect, and the endowment measurement error is $\nu_i = u_i - e_i\hat{\Gamma}$.

If there are no other observables or unobservables, the estimated (linear) calorie allocation equation is

$$(16) \quad C_i = a + b\hat{\mu}_i + e_i.$$

The least-squares estimator of b , if the true health endowments are nonstochastic and $(1/n)\sum \mu_i^2$ converges as $n \rightarrow \infty$ to a positive finite limit $\sigma_{\mu\mu}$, is

$$(17) \quad \text{plim}_{n \rightarrow \infty} \hat{b} = b \frac{\sigma_{\mu\mu}}{\sigma_{\mu\mu} + \sigma_{\nu\nu}} + \frac{\sigma_{e\nu}}{\sigma_{\mu\mu} + \sigma_{\nu\nu}}.$$

The first term in (17) corresponds to the classical error-in-variables bias. However, the second term appears because of the indirect estimation of the endowment from (12). When calorie consumption has a positive marginal product in the production of health ($\hat{\Gamma} > 0$), then from (15), $\sigma_{e\nu} < 0$. Thus, if b is positive, \hat{b} will underestimate b unambiguously, but if the true endowment effect is negative, the sign of the bias is indeterminate. The (biased) least-squares estimator of the error-ridden endowment effect would tend to reject reinforcement with respect to calories if it in fact were true; errors in measurement in calories will make households appear to be more com-

pensatory with respect to calories than they really are.⁸

Consistent parameter estimates in the presence of errors in variables can be obtained by using instrumental-variables methods. Repeated observations on the measured-with-error variable or the availability of different but related indicators of the phenomena to be measured (for example, multiple-proxy variables) are potential sources of instruments. If individual-specific food intake and anthropometric measures of health were measured at more than one point in time, then even if all measurements of (calorie) consumption are made with error, all that is required for consistent instrumental-variables estimation is that the period-specific errors be uncorrelated across time periods. Moreover, if there are available multiple indicators of health and thus of endowments in addition to repeated measures, they can be used as well, as long as the measurement errors in each endowment type are uncorrelated across time periods (noncontemporaneously) with the health-endowment measurement error.

III. Empirical Results

A. Data and the Estimation of the Health Technology

As the previous discussion has made clear, to obtain direct estimates of endowment effects requires data that not only provide individual-specific information on health and consumption but contain i) sufficient cross-sectional variation in exogenous variables needed as instruments for estimation of the health technology and (ii) repeated observations on individuals to purge estimated endowments of measurement errors. The 1981-2 Nutrition Survey of Rural Bangladesh (Kamaluddin Ahmad and Hassan, 1986) provides information on individual-specific food consumption and anthro-

⁸The parameters associated with all other regressors measured without error are also biased; the sign of their bias can be determined from the variance-covariance matrix of the observations (Maurice Levi, 1973).

pometric measures of health along with other individual and household attributes for 385 households in 15 villages scattered throughout Bangladesh.⁹ Intrahousehold food-consumption information was collected once for 25 (out of 50 sampled) households in each of 12 randomly selected villages and in 35 (out of 70 sampled) households in an industrial town. In addition, the same information was collected at four separate times within a year for 25 (out of 50 sampled) households in each of two of the remaining villages. These Bangladesh data thus permit estimation of the health technology from the cross-sectional sample, as well as estimation of endowment responses purged of measurement error, based on the longitudinal component of the data set.¹⁰

The intrahousehold dietary information was collected by specially trained female dietary investigators who measured dietary intake by weighing each individual's intake in the home over a 24-hour period. All individuals covered by the dietary survey were also examined by a clinician, who obtained measures of weight, height, skinfold thickness, and mid-arm circumference. Information was also obtained on the occupation of each household member, and the energy intensity of his or her activity was coded using guidelines established by the Food and Agriculture Organization and the World Health Organization (see Appendix A). The prices of a wide variety of foods sold in the village market were separately obtained in the survey, so that there is one price per commodity per village.

We use the information on weight-for-height to measure health, which is considered a good short-run measure of nutri-

tional status that will be sensitive to daily food consumption and activity levels. To estimate the health technology (1), food consumption was converted into nutrient intakes using conversion factors specific to Bangladeshi foods (Institute of Nutrition and Food Science, 1980). Calorie consumption, however, would appear to be a sufficient indicator of nutritional intake. The typical Bangladeshi diet is very simple; cereals account for 87 percent of calorie consumption, as well as 78, 82, 84, 70, and 82 percent of the consumption of protein, iron, thiamine, riboflavin, and niacin, respectively. As the consumption of each nutrient is a linear function of all foods consumed, the large share of consumption derived from just one food group makes the set of observed nutrient intakes nearly perfectly collinear. Moreover, we would expect that weight-for-height, as an indicator of short-run health, should not respond substantially if at all to the intake of any nutrient except calories. Daily changes in weight reflect the difference between calories consumed and calories expended.

To reflect calorie outflow, we add to individual-specific nutrient consumption in the weight-for-height production function two dummy variables reflecting participation in occupations categorized as "very active" or "exceptionally active" based on the occupational data.¹¹ We add as well dummy variables indicating whether a woman was pregnant or lactating at the time of the survey. Exogenous regressors included are age, age squared, sex, the interaction of sex and age, and a set of dummy variables indicating the source of the household's drinking water (well, pond, tube well, or river/canal). In

⁹The data from one additional village of hill tribes (who are not racially or ethnically related to Bengalis) were not used in our analysis, as their dietary and other behaviors are considered too unlike those of ethnic Bengalis.

¹⁰Hassan (1984) has compared the nutritional information in the survey with that collected in prior nutrition surveys in Bangladesh (and East Pakistan) to draw inferences concerning trends in Bangladeshi health and food consumption.

¹¹A possibly superior procedure would have been to employ individual dummy variables for each of the 14 occupations provided in the data. Because we treat occupation as a choice variable, however, we would need more instruments than we have available to identify all of the individual activity effects on weight-for-height. The FAO/WHO categories provide a parsimonious way of representing occupations in terms of their consequences for short-term health. If the categorization is correct, our estimates are more efficient than those that would be obtained from the more agnostic specification, if it could be estimated.

accord with the model, we treat nutrients, activities, pregnancy, and lactation as endogenous variables and estimate the production function using two-stage least squares. Identifying instruments are the household head's age and schooling level, household landholdings, and the village food prices interacted with household landholdings, the head's schooling and age, and the individual age and sex variables. The food prices are those for rice, wheat flour, potatoes, leafy vegetables, okra, green chilies, sugar and sweets, eggs, mustard oil, pulses, fish, milk, onions, garlic, and meat.¹²

Table 3 presents both (inconsistent) ordinary least squares (OLS) and consistent two-stage least squares (2SLS) estimates of the parameters of the Cobb-Douglas production function for weight-for-height. These estimates are obtained using the cross-sectional component of the data describing the full set of 15 villages (with one round from each of the two multiple-round villages). The calorie elasticity is seriously underestimated by OLS, although it is positive and statistically significant using either procedure. Moreover, the OLS estimates of the effect of the energy intensity of effort on weight-for-height are of the opposite sign to the consistent 2SLS estimates, indicating a possible strong relationship between activity choice and the health residual, containing the endowment. The 2SLS estimates indicate that increased calorie consumption significantly increases weight-for-height and demonstrate that participation in exceptionally active occupations tends to deplete weight-for-height, although the estimated activity coefficient has a relatively large standard error. The less active occupations categorized as "very active" have an estimated coefficient only one-eighth that of "exceptionally active" occupations.

We also tested whether calorie consumption was a sufficient statistic for nutrient consumption and whether the calorie elasticity differed between males and females. We could not reject the null hypothesis that

TABLE 3—EFFECTS OF CALORIE CONSUMPTION, ACTIVITY LEVEL, AND PREGNANCY STATUS ON WEIGHT-FOR-HEIGHT

Variable ^a	Ordinary least-squares estimates ^c	Two-Stage least-squares estimates ^c
Calorie consumption ^b	0.0295 (4.09)	0.136 (3.37)
Very active occupation ^b	0.0859 (5.34)	-0.0119 (0.23)
Exceptionally active occupation ^b	0.0668 (3.43)	-0.0817 (1.26)
Pregnant ^b	0.262 (7.69)	0.326 (1.34)
Lactating ^b	0.144 (9.28)	0.513 (4.65)
Age	0.284 (16.6)	0.0987 (1.90)
Age squared	-0.00456 (1.44)	0.0174 (2.37)
Sex (male = 1)	0.00196 (0.08)	-0.0578 (1.81)
Age × sex	0.0152 (1.74)	0.0687 (4.04)
Water drawn from tube well	-0.0478 (3.13)	-0.0406 (2.10)
Water drawn from well	-0.0720 (4.11)	-0.0693 (3.15)
Water drawn from pond	-0.0460 (2.30)	-0.0649 (2.55)
Constant	-2.56 (52.4)	-3.12 (13.9)
<i>N</i>	1,737	1,737
<i>R</i> ²	0.775	—
<i>F</i>	395.1	—
<i>H</i> ₀ : No influence of calcium, carotene, thiamine, and riboflavin consumption ^b (<i>F</i>)	—	1.23
<i>H</i> ₀ : No difference in effect of calorie consumption by sex (<i>F</i>)	—	2.16

^aAll variables in logs, except sex, water sources, and activity level.

^bEndogenous variable; instruments include household head's age and schooling level, landholdings, and prices of all foods consumed interacted with individual age and sex variables, land, and head's schooling and age.

^cAsymptotic *t* ratios in parentheses.

four additional nutrients found by James Ryan et al. (1984) to be potentially important determinants of short-run health in a rural area of India—calcium, carotene, thiamine, and riboflavin—do not influence

¹²Pitt (1983) shows that nutrient consumption is significantly responsive to food prices in Bangladesh.

weight-for-height in our sample. ($F_{[4, 1724]} = 1.23$). The null hypothesis that the calorie-output elasticity is the same for men and women also could not be rejected ($F_{[1, 1724]} = 2.16$). Thus, the difference between sexes in the production of weight-for-height is sufficiently well specified as an age-dependent intercept shift. Except for the first 2.5 years of life, Bangladeshi males are predicted to have greater weight-for-height than females having identical levels of inputs.

B. Endowments and Calorie Consumption

Having obtained estimates of the health technology, we can compute health (weight-for-height) endowments for each individual based on actual calorie consumption and activity. In order to use the repeated-measure methodology to mitigate the effects of errors in measurement, we use the longitudinal component of the sample, which provides four rounds of data for 50 households in two of the villages (Jorbaria and Falshatia). We also use as instruments estimated endowments of mid-arm circumference and skinfold thickness derived from production functions, estimated by two-stage least squares, containing the same regressors and instruments as the weight-for-height production function. Three instruments for an individual's weight-for-height endowment in a period τ are thus constructed: the estimated endowments of the three health attributes averaged over the survey rounds in which the individual was present *excluding* period τ .¹³

¹³Formally, the set of instruments $Z_{i\tau}^j$ associated with the endowment of type j for individual i in period τ is constructed as

$$Z_{i\tau}^j = \frac{1}{T_i - 1} \sum_{t \neq \tau} \hat{\mu}_{it}^j$$

where j = weight-for-height, skinfold thickness, or mid-arm circumference, and where T_i is the number of repeated measures (rounds) available for person i . Instruments for the mean weight-for-height endowments of groups (classes) of family members in period τ are constructed as the group means of the individual-specific means $Z_{i\tau}^j$.

The household welfare-maximization model, as noted, implies that the calories allocated to an individual in the household depend on that person's characteristics (age, sex, and endowment), the characteristics of all other household members, and household or village-specific characteristics such as health-program availability and food prices. With respect to village-level variables, because our longitudinal sample is taken from only two villages, a village dummy variable captures all village-specific determinants. To summarize parsimoniously the intrahousehold distribution of the exogenous characteristics of household members, we computed the household means of those variables, namely mean age, mean age squared (variance of ages), proportion of household members male, and the mean of the household's endowments. Household-specific variables include water sources and family income. Family income is treated as an endogenous variable because wages are assumed to depend on endowments, calorie allocations, and the level of effort.

The first column of Table 4 provides two-stage generalized least-squares estimates of the logarithmic calorie-allocation equation, estimated with the full sample of individuals from the two villages, but in which instruments are not used for the endowment variables. The second column provides parameter estimates that use the instruments for the endowment variables. As predicted, the uninstrumented coefficient estimate for own endowment is algebraically less than the (positively signed) instrumented own endowment coefficient estimate. Indeed, it is of opposite sign, indicating compensation when there is evidently net reinforcement with respect to calories.¹⁴ The uninstrumented family or cross-endow-

¹⁴The coefficient on own endowment in these logarithmic calorie-allocation equations should be interpreted as the elasticity of own health with respect to own endowment conditional on mean family endowments remaining fixed. This elasticity then corresponds to the experiment in which a transfer of endowment occurs within the household that leaves mean endowments unchanged. This same interpretation also applies to the own age and sex coefficients.

TABLE 4—TWO-STAGE GENERALIZED LEAST-SQUARES ESTIMATES:
EFFECTS OF PERSONAL AND FAMILY CHARACTERISTICS ON THE
ALLOCATION OF PERSONAL CALORIE CONSUMPTION

Variable ^a	Two-stage least-squares estimates ^b			
	All family members		By sex	
	No instruments for endowments	Instruments for endowments ^c	Males	Females
Individual weight/height endowment	-0.145 (1.98)	0.132 (1.33)	0.676 (4.14)	0.0662 (0.26)
Family endowment	-0.867 (1.01)	-1.15 (0.75)	—	—
Family endowment, males	—	—	-0.743 (1.72)	-0.414 (1.27)
Family endowment, females	—	—	-0.325 (1.10)	-0.0709 (0.20)
Family income ^{c,d}	0.0640 (2.04)	0.122 (3.95)	0.0839 (1.98)	0.0961 (2.07)
Age	1.35 (25.8)	1.35 (24.7)	1.44 (17.5)	1.33 (15.7)
Age squared	-0.199 (19.8)	-0.199 (19.0)	-0.200 (12.9)	-0.196 (11.6)
Sex (male = 1)	-0.0191 (0.24)	0.0492 (0.60)	—	—
Age × sex	-0.0679 (2.63)	0.0801 (2.97)	—	—
Mean age of family members	-0.0711 (1.10)	-0.114 (1.66)	0.00318 (0.03)	-0.122 (1.44)
Variance of ages of family members	-0.0757 (1.39)	-0.115 (2.03)	-0.0837 (0.99)	-0.120 (1.61)
Proportion of family members male	-0.0330 (0.32)	-0.0588 (0.54)	-0.00762 (0.04)	-0.0777 (0.55)
Water drawn from tube well	0.227 (3.03)	0.221 (2.81)	0.162 (1.58)	0.271 (3.29)
Jorbaria village	0.245 (3.36)	0.254 (3.34)	0.196 (1.95)	0.283 (3.54)
Constant	4.86 (18.3)	4.65 (16.7)	4.52 (10.3)	4.82 (13.2)
<i>N</i>	806	806	407	371
<i>X</i> ² (no family error component) ^e	245.6	243.5	129.0	38.06
Share of family error component variance in total error variance	0.205	0.204	0.234	0.218

^aAll variables in logs, except sex, water source, location, and sex ratio.

^bAsymptotic *t* ratios in parentheses.

^cInstruments for income and endowments are: household landholdings and household head's schooling and age; and means of individual and family endowments for weight/height, skinfold thickness, and arm circumference calculated over all survey rounds, excluding the round from which observation is drawn.

^dEndogenous variable.

^eLagrange multiplier (Breusch-Pagan test).

ment parameter is algebraically greater than the consistent estimate, but as the consistent estimate is negative, the sign of the bias could not be predicted unambiguously *a priori*. A comparison of the first two columns in Table 4 also reveals that other parameters are biased substantially as well.

The consistent estimate of the own endowment effect in column 2 of Table 4 suggests that there is reinforcement with respect to calories, although the coefficient is not statistically different from zero at standard levels of significance ($t = 1.32$). However, the activity distributions reported in Table 2 indicate that there are important gender differences in activities, and thus, as our framework suggests, endowment effects may differ by gender. In columns 3 and 4 of Table 4, we provide two-stage generalized least-squares estimates of calorie-allocation equations stratified by sex (with all endowments instrumented).¹⁵ The estimates indicate that a 10-percent increase in a male's endowment increases his calorie allocation by 6.8 percent; the own endowment effect for females is one-tenth that of males. These differences are consistent with the theory, given the lack of participation by women in energy-intensive activities and the findings in similar settings that health matters for the wages of men but not women.

Corresponding to the positive and significant own endowment effect for males, the cross effect of the endowment of other males in the household is negative. These results thus reject the pure income-maximizing model, since if households allocate calories and effort so as to maximize income, all cross effects will be zero. The theory also predicts that, if there is calorie reinforcement, the effect of an increase in a female's health endowment on the calories allocated to others in the household should be less in absolute value than the effect of an increase in a male's health endowment, if health status is less important for women in their

activities. In both the male and female calorie-allocation equations of Table 4 (columns 3 and 4), the effect of the mean endowment of females on calorie consumption is indeed considerably less in absolute value than the effect of the mean endowment of males, although the difference is not statistically significant because of the imprecision with which both endowment effects are measured.

The parameter estimates reported in Table 4 may specify cross effects in an imperfect manner by assuming that they can be represented by the means (and higher moments) of household distributions, although adding second- and third-order moments did not significantly improve the fit of these equations. A household fixed-effects estimator, however, provides an estimate of own endowment effects that requires no assumptions about the parameterization of the household variables, because the full set of cross terms and household-specific regressors are impounded in the fixed effect. Although this approach is more likely to avoid specification error, it, of course, prevents identification of the parameters associated with family endowments and other household-specific regressors.

Table 5 reports household fixed-effects two-stage generalized least-squares estimates, by gender, of the effects of personal characteristics on individual calorie consumption. The set of instruments for the own endowment measure remains the same. The estimation procedure takes into account the sample property that individuals appear more than once. The null hypothesis of no individual-specific error components is indeed rejected by the Breusch-Pagan (Lagrange multiplier) test statistic in each equation estimated.¹⁶

¹⁵The reduction in total sample size that occurs when the sample is stratified by sex results from the necessity of using only those households that have both males and females in order to estimate gender-specific cross and own endowment effects.

¹⁶Note that only household (and not individual) random effects were specified in estimating the calorie-allocation equations of Table 4. The Breusch-Pagan (Trevor S. Breusch and Adrian Pagan, 1980) statistics of Table 5 confirm the importance of individual effects even when controlling for household fixed effects. The parameter estimates of Table 4 are nonetheless consistent, but standard errors are underestimated by about 10 percent, based on our experience in obtaining the estimates reported in Table 5.

TABLE 5—FIXED-EFFECTS TWO-STAGE GENERALIZED LEAST SQUARES:
EFFECTS OF PERSONAL CHARACTERISTICS ON
INDIVIDUAL CALORIE CONSUMPTION

Variable ^a	Two-stage least-squares estimates ^b			
	Males		Females	
	Endowment effects constant	Endowment effects vary with age	Endowment effects constant	Endowment effects vary with age
Own endowment ^c	0.447 (3.58)	—	-0.0278 (0.15)	—
Age < 6 ^c	—	-0.435 (1.35)	—	-0.314 (0.46)
6 ≤ age < 12 ^c	—	0.923 (2.29)	—	1.86 (2.13)
Age ≥ 12 ^c	—	1.21 (2.69)	—	0.0894 (0.13)
Age	1.44 (22.9)	1.31 (14.9)	1.34 (18.1)	1.35 (17.9)
Age squared	-0.201 (16.7)	-0.170 (9.16)	-0.199 (13.4)	-0.206 (13.7)
<i>N</i>	429	429	371	371
<i>X</i> ² (no individual error components)	46.5	48.35	32.36	26.17
Individual error variance/total error variance	0.287	0.300	0.258	0.282

^aAll variables in logs.

^bAsymptotic *t* ratios in parentheses.

^cInstrumental variables used are means of individual and family endowments for weight/height, skinfold thickness, and arm circumference calculated over all survey rounds, excluding the round from which observation is drawn.

Columns 1 and 3 of Table 5 report the within-household, gender-specific (logarithmic) calorie-allocation equations having own endowment, own age, and own age squared as regressors. The parameter estimates diverge very little from those reported in Table 4. The elasticity of calorie consumption with respect to own health endowment is 0.447 ($t = 3.58$) for males, indicating reinforcement, and is only -0.028 ($t = -0.15$) for females.

In columns 2 and 4 of Table 5, we report estimates obtained using specifications of the calorie equation in which sex-specific own endowment effects are allowed to vary across the three age groups that appear from Table 2 to be related to the differentiation of activity patterns. The pattern of estimated own endowment effects matches up well with the pattern of activities presented in Table 2. Both male and female

young children (aged less than six years) have the (algebraically) smallest own endowment effects. There is no labor-market return to higher endowments for these family members, and thus calorie compensation dominates; part of the better health derived from a higher endowment is "taxed" away by the household, in this case solely via the allocation of foods.

Male and female children aged 6–12 years evidently engage in a more diverse set of activities ranked by energy intensity, and the own endowment parameters exhibit reinforcement and are statistically significant for both males and females. A 10-percent increase in the health endowment of a 6–12-year-old child increases calorie consumption by 9.2 percent if the child is male and 18.6 percent if the child is female. The higher rate of reinforcement for girls in this age group is consistent with their greater

diversity of activities, categorized by energy intensity, displayed in Table 2. Adult males exhibit the greatest diversity of activity choice ranked by energy intensity among all age-sex groups, while adult females have very limited diversity and are concentrated in less energy-intensive activities. Reflecting this, the estimated own endowment elasticity of calorie consumption for adult males is positive, statistically significant, and the largest of all groups (1.21), while that for adult females is close to zero (0.09).

If β_k is the estimated own endowment effect for age-sex group k , then the variability in consumption in group k depends on β_k and on the group's dispersion in endowments, as $\text{Var}(c^k) = \beta_k^2 \text{Var}(\mu^k)$. Based on our estimates of the endowments, we cannot reject the hypothesis that all within-group endowment variances are equal. Our estimates of β_k values thus imply that household allocation rules and the variation in the effects of health on productivity across activities are in part responsible for the higher variability in intrahousehold calorie allocations among males relative to females for individuals aged 12 and over. These factors also contribute to the variability among girls and boys for those household members aged between 6 and 12 years, with no effect among boys and girls less than 6.¹⁷

C. Endowments, Family Income, and Activity Participation: Household Discrimination

Although the results thus far are consistent with there being a return to health in the labor market, it remains to demonstrate with these data that income is positively associated with health endowments and that individuals with higher endowments are more likely to choose activities with a greater

energy intensity of effort, as implied by the theory. Moreover, with estimates of endowment effects on energy intensity, we can use (11) to compute how the household on net responds to differences in endowments for each gender. While the Bangladesh data do not provide information on individual-specific wage rates or earnings, we can test whether households with higher average health endowments among adult males have higher incomes, given land resources and schooling. Table 6 (column 1) provides estimates of the determinants of (log) per capita income. These show that income is positively and significantly associated with the average endowments of males older than 12 years of age, but as expected, the adult female endowment elasticity of income is only one-sixth as large as the male endowment elasticity and not statistically different from zero.

Table 6 (column 2) also reports maximum-likelihood (ML) instrumental-variable estimates of a probit activity-choice equation for individuals aged 12–60.¹⁸ The dichotomous dependent variable in this equation has the value of 1 if an adult is engaged in an exceptionally active occupation (the only activity category that substantially reduced weight-for-height in the estimated production function; Table 3) and 0 otherwise. Here, own endowment has a positive and statistically significant (at the 10-percent level) effect on the probability of participating in an exceptionally active activity. In addition, consistent with the calorie-allocation estimates of Table 4, the male family endowment has a large negative influence on this probability, five times larger than the influence of the female family endowment. The coefficient on sex (male = 1) is positive and statistically significant, reflecting the differences between sexes in the diversity of occupations, given endowments. Thus, the results reported in Table 6 confirm that there is a pecuniary return to health and effort, that adult males with

¹⁷To test for seasonality in endowment effects, we tested whether endowment responses varied with income by interacting the endowment and age variables with household income using the household fixed-effects procedure. We could not reject the hypothesis that endowment and age effects were independent of income levels.

¹⁸The likelihood maximized is given in Richard J. Smith and Richard W. Blundell (1986).

TABLE 6—DETERMINANTS OF THE LOG OF PER CAPITA HOUSEHOLD INCOME AND PROBABILITY OF PARTICIPATING IN AN "EXCEPTIONALLY ACTIVE" OCCUPATION AMONG PERSONS AGED 12–60 YEARS

Variable	Per capita income (two-stage least-squares) ^a	Exceptionally active occupation (full-information ML IV probit) ^a
Own endowment ^b	—	13.9 (1.64)
Family endowment	2.38	–16.74
males ≥ 12 years old ^b	(2.86) ^b	(2.29)
Family endowment	0.378	–3.67
females ≥ 12 years old ^b	(0.75)	(1.25)
Age	—	1.28 (1.06)
Sex	—	6.92 (2.36)
Landholding	0.0200 (0.64)	–0.0219 (2.46)
Household head's schooling	0.109 (1.80)	–1.09 (1.54)
Mean age of family members	0.0444 (0.14)	–1.58 (1.18)
Variance of ages of family members	0.591 (1.91)	–5.55 (2.50)
Proportion of family members male	0.566 (1.09)	–4.11 (1.82)
Jorbaria village	–0.199 (1.30)	–5.98 (2.19)
Constant	4.23 (4.11)	4.95 (1.17)
<i>N</i>	45	153
<i>F</i>	3.73	—
$X^2_{[12]}$	—	76.2

^aAsymptotic *t* ratios in parentheses.

^bInstrumented.

higher endowments are more likely to undertake exceptionally energy-intensive work, and that adult female health endowments are relatively unimportant in determining activity choices or household income compared to adult male endowments.

Finally, the net effect of a change in own endowment on own health [eq. (11)] can be calculated from the estimates of the health technology in Table 3 and the estimated endowment effects on calories and activities in Tables 5 and 6. For both adult males and females (aged 12 years and above), our estimates indicate that, in addition to its direct effect on health, an increase in endowment tends to increase health by increasing calorie consumption and to reduce health by

inducing greater intensity of effort. The latter indirect effect dominates the former for both sexes. The elasticity of own health with respect to own endowment is 0.88 for adult males and 0.97 for adult females.¹⁹

¹⁹The elasticity is given by

$$d \ln h / d \ln \mu = 1 + (\partial \ln h / \partial \ln c)(d \ln c / d \ln \mu) \\ + (\partial \ln h / \partial e)(de / d \ln \mu).$$

The health elasticity of calories, the first parenthetical term in the elasticity expression, is 0.136 for both males and females (Table 1). The elasticity of calorie consumption with respect to own health endowment ($d \ln c / d \ln \mu$) is 1.21 for adult males and 0.089 for

Bangladesh households thus exhibit compensatory behavior with respect to health. Moreover, as the difference between the endowment elasticity and unity can be thought of as a "tax" levied by the household on the exogenous health of its members, our estimates indicate that the exogenous health of adult males is taxed at a higher rate than the exogenous health of adult females (12 percent vs. 3 percent).

IV. Conclusion

In this paper, we have examined the determinants of calorie consumption and activity choices from the perspective of a model of intrahousehold allocation that incorporates individual heterogeneity in exogenous healthiness and differences in labor-market returns to health and effort across groups of individuals. The empirical analysis was applied to individual and household-level data from Bangladesh, a country that exhibits large differences in calorie consumption and in the energy intensity of activity by age and gender. Our results reveal that energy-intensive effort tends to reduce health as measured by weight-for-height, that there is a pecuniary return to health and effort, and that there is substantial calorie reinforcement for those classes of individuals best able to alter the energy intensity of effort. In particular, adult males (aged 12 years and above) and male and female children (aged 6–12) were found to receive calorie reinforcement with respect to their health endowments. These classes of individuals were also those exhibiting the most diverse activity choices ranked by energy intensity. Thus, linkages

between health levels and productivity, combined with the circumscribed activities of adult women in Bangladesh, appear to account for part of the disparities in the average consumption of nutrients across adult men and women and to contribute to the greater variability among men in nutrient consumption.

Our results also reject the income-maximizing model of the household in favor of a model in which households exhibit some aversion to inequality. Indeed, even though the rate of calorie reinforcement for adult males was quite high (1.21 in elasticity) and almost zero for adult females, the greater likelihood of adult males with higher endowments to undertake exceptionally energy-intensive work resulted in a "tax" on adult male endowments that exceeded that of adult females (12 percent vs. 3 percent), signaling some discrimination against males by the household.

Our evidence that disparities by gender in food consumption in a low-income society like Bangladesh reflect the gender differentiation in the energy intensities of activities suggests that increases in labor-force opportunities for women, *ceteris paribus*, will likely increase the calories allocated to women. However, as we have shown, the health and welfare benefits of such an increase in calorie consumption by women will be tempered by the increased level of energy-intensive activity associated with greater calorie consumption. Furthermore, while an increase in the occupational diversity of women is likely to reduce (calorie) consumption inequality between the sexes, it will increase inequality among adult women and thus may increase overall inequality in consumption and health. The increase in inequality among women will reflect the increased importance of the distribution of endowments in determining the distribution of calories when there is a greater return to effort and health in the labor market. However, to the extent that economic development is characterized by a transformation of work activities to those in which linkages between food consumption and productivity are weak, overall inequality in food consumption may be attenuated for all groups as incomes rise.

adult females (Table 5). The effect of participation in an exceptionally active occupation on (log) health ($\partial \ln h / \partial e$) is -0.082 for both males and females (Table 3). The estimated effect of (log) own health endowment on the probability of participating in an exceptionally active occupation is 13.9 (Table 6) multiplied by the standard normal density function evaluated at the value of the underlying latent activity variables for adult males and females, which are 0.247 and 0.038, respectively.

REFERENCES

- Ahmad, Kamaluddin and Hassan, Md. Nazmul, *Nutrition Survey of Rural Bangladesh 1981-82*, Dhaka: Institute of Food Science and Nutrition, 1986.
- Bardhan, Pranab K., "On Life and Death Questions," *Economic and Political Weekly*, 10-24 August 1974, 9, 1293-1305.
- Behrman, Jere, "Intrahousehold Allocation of Nutrients in Rural India: Are Boys Favored? Do Parents Exhibit Inequality Aversion?" *Oxford Economic Papers*, March 1988, New Series, 40, 32-54.
- _____, "Intrahousehold Allocation of Nutrients and Gender Effects: A Survey of Structural and Reduced Form Estimates," in Siddig R. Osmiani, ed., *Nutrition and Poverty*, Oxford: Oxford University Press, 1990, forthcoming.
- _____, and Deolalikar, Anil, "Agricultural Wages in India: The Role of Health, Nutrition and Seasonality," in David E. Sahn, ed., *Seasonal Variability in Third World Agriculture. The Consequences for Food Security*, Baltimore: Johns Hopkins University Press, 1989, 107-17.
- Breusch, Trevor S. and Pagan, Adrian R., "The Lagrange Multiplier Test and its Applications to Model Specification' in Econometrics," *Review of Economic Studies*, September 1980, 47, 239-53.
- Dasgupta, Partha and Ray, Debraj, "Inequality, Malnutrition, and Unemployment: A Critique of the Market Mechanism," mimeo, Department of Economics, Stanford University, 1984.
- Deolalikar, Anil B., "Do Health and Nutrition Influence Labor Productivity in Agriculture? Econometric Estimates for Rural India," *Review of Economics and Statistics*, August 1988, 70, 406-13.
- Gersovitz, Mark, "Savings and Nutrition at Low Incomes," *Journal of Political Economy*, October 1983, 91, 841-55.
- Hassan, Md. Nazmul, "Studies on Food and Nutrient Intake by Rural Population of Bangladesh: Comparison Between Intakes of 1962-64, 1975-76, and 1981-82." *Ecology of Food and Nutrition*, September 1984, 15, 143-58.
- Leibenstein, Harvey A., *Economic Backwardness and Economic Growth*, New York: Wiley, 1957.
- Levi, Maurice D., "Errors in Variables Bias in the Presence of Correctly Measured Variables," *Econometrica*, July 1973, 41, 985-6.
- Pitt, Mark M., "Food Preferences and Nutrition in Rural Bangladesh," *Review of Economics and Statistics*, February 1983, 65, 105-14.
- Rosenzweig, Mark R. and Schultz, T. Paul, "Market Opportunities, Genetic Endowments, and Intrafamily Resource Distribution: Child Survival in Rural India," *American Economic Review*, September 1982, 72, 803-15.
- _____, and _____, "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birthweight," *Journal of Political Economy*, October 1983, 91, 723-46.
- Ryan, James G., Bodinger, P. D., Rao, N. P. and Pushpamma, P., *Determinants of Individual Diets and Nutritional Status in Six Villages of South India*, Hyderabad: International Crop Research Institute for the Semi-Arid Tropics, 1984.
- Sahn, David and Alderman, Harold, "The Effect of Human Capital on Wages, and the Determinants of Labor Supply in a Developing Country," *Journal of Development Economics*, September 1988, 29, 157-83.
- Smith, Richard J. and Blundell, Richard W., "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, May 1986, 54, 679-86.
- Stiglitz, Joseph E., "The Efficiency Wage Hypothesis, Surplus Labor, and the Distribution of Income in LDC's," *Oxford Economic Papers*, May 1976, New Series, 28, 185-207.
- Strauss, John, "Does Better Nutrition Raise Farm Productivity?" *Journal of Political Economy*, April 1986, 94, 297-320.
- Institute of Food Science and Nutrition, University of Dhaka, *Nutritive Values of Local Foodstuffs*, Dhaka: University of Dhaka, 1980.

A Social Exchange Approach to Voluntary Cooperation

By HEINZ HOLLÄNDER*

A social exchange approach to voluntary cooperation is developed on the assumption that voluntary cooperative behavior is motivated by social approval, which is conceptualized as an emotional activity. The associated unique Nash equilibrium may have attractive welfare properties and provides an understanding of spontaneous norm emergence. Furthermore, the opening of a market or government intervention for the collective good is shown to affect voluntary cooperation negatively. (JEL 024, 025)

This article deals with the problem of voluntary cooperation for a pure collective good when the impact of feasible individual contributions on its supply can be neglected. According to standard theory, rational agents will entirely fail to cooperate because they are caught in a Prisoner's Dilemma. However, as many have observed, the empirical extent of cooperation is in marked contrast to the one expected from standard theory.¹ Thus, there is need for a more sophisticated theory that attributes success or failure of cooperation to the circumstances at hand. As is well known, several explanations of voluntary cooperation have been put forward,² but they seem to

be exposed to severe criticism,³ or, as Olson's (1965) by-product theory, can at best account for a small part of observed cooperative behavior. This paper offers a simple axiomatic model of social exchange in which cooperative behavior is motivated by the expectation of emotionally prompted social approval and explores some of its implications. The basic hypothesis has a long tradition. It was employed, for instance, by Bernhard de Mandeville (1714) in his *Fable of the Bees*, by Adam Smith (1759) in his *Theory of Moral Sentiments*, and in modern sociology, by George Caspar Homans (1961) in his *Social Behavior*. In order to demonstrate the logic of the social exchange approach as clearly as possible, I have chosen simple and restrictive but, I hope, nevertheless plausible assumptions supporting the main results. The contribution is intended to be an exemplary argument rather than a general theory.

In the model, individual cooperative contributions and, thus, the supply of the collective good depend on the strength of the approval incentive, while the latter depends on individual contributions and the supply of the collective good. I show the existence of a unique symmetric Nash equilibrium and give a full characterization of the equilibrium solution. Furthermore, it is shown that the model provides a micro-founded

*Department of Economics and Social Sciences, University of Dortmund, 4600 Dortmund 50, Federal Republic of Germany. I thank Matthias Fischer, Jürgen Frank, Franz Haslinger, Wolfram Richter, and Joachim Weimann for discussion and/or helpful suggestions. I am also grateful to three anonymous referees for their comments.

¹For instance, people join political parties, trade unions, the Red Cross, and other nonprofit organizations; they turn out for elections, demonstrations, and strikes; many give considerable amounts to charity and make voluntary blood donations; and, finally, they often help others, show consideration for others, line up, are honest, don't cheat whenever they can get away with it, and comply with other behavioral norms.

²The more important ones are the by-product approach of Mancur Olson (1965); the altruism approach of, for instance, Robert Schwartz (1970), Gary S. Becker (1974), David A. Collard (1978 Ch. 10), Kenneth J. Arrow (1981), and Howard Margolis (1982); the iterated game approach of, for instance, Peter Hammond (1975), Mordecai Kurz (1977), and Andrew Schotter (1981); and the sociobiological approach of, for instance, Edward O. Wilson (1975) and Robert Axelrod (1984).

³For instance, the altruism approach has been criticized by Robert Sugden (1982, 1984), the iterated-game approach has been criticized by Dennis C. Mueller (1986), and the sociobiological approach has been criticized by Becker (1976) and Philip Kitcher (1985).

invisible-hand explanation of the emergence of a behavioral norm. In general, the social exchange allocation is not Pareto-efficient. However, compared to the optimal planning allocation and the hypothetically ideal market allocation without approval incentives, it may provide more of the collective good and higher group welfare. A further result confirms the argument of Fred Hirsch (1976) that the opening of a market for the collective good supersedes voluntary contributions, at least partly, and possibly decreases social welfare.

In Section I, the model setting is developed. Section II contains a formal description of rational and emotional behavior, and it explores the consequences of consistent social interaction. The relationship between the sociological concept of a behavioral norm and the social exchange equilibrium is discussed in Section III. The welfare analysis is done in Section IV, and Section V offers a few concluding remarks.

I. The Model Setting

The model proceeds from a group of n identical agents. Initially, every agent is provided with π units of a private good. If b units are contributed to the collective good, the remaining units, p , are privately consumed:

$$(1) \quad p = \pi - b, \quad 0 \leq b \leq \pi.$$

The collective good is produced by means of aggregate forgone private consumption alone. For simplicity, I assume constant returns to scale with costs proportional to group size. Thus, the amount produced is

$$(2) \quad c = \frac{1}{n} \sum_{i=1}^n b_i.$$

Furthermore, it is assumed that the group is large in the sense that the effect of every feasible individual contribution on collective good supply can be neglected.

A cooperative contribution satisfies the following three conditions that characterize

a pure gift: first, it confers some benefit upon others; second, it imposes net costs upon the cooperative agent; and finally, it is voluntary. This suggests that other group members will behave toward cooperative agents in essentially the same way as donees do toward donors. In general at least, recipients of gifts react with, for instance, gratitude or sympathy for their benefactors. These reactions typically are not rationally calculated but, rather, prompted by the stimulus-response mechanism of the human emotional system. In analogy to the gift case and in accordance with Homans (1961), I proceed on the assumption that some subjective value of a cooperative contribution b as assessed by the reacting agent measures the stimulus power $s(b)$ prompting emotional reactions. It should be noted that the introduction of emotional activities prompted by some stimulus takes us beyond rationally calculating economic man.

Two categories of emotional reactions can be distinguished. The first comprises emotions, feelings, and attitudes as inner states. To the second belong all activities associated with these inner states such as, for instance, facial expressions, verbal expressions of gratitude, and even killing in the heat of passion. These activities can be regarded as expressions of inner states. In accordance with Homans (1961) and Smith (1759), an emotional reaction consisting of an inner state and its particular expression is called a sentiment.

There seem to be quite a number of different classes of sentiments respectively characterized by the extremes sympathy and antipathy, love and hate, gratitude and resentment, joy and grief, pride and shame, admiration and contempt, and presumably some others. Within each class, the extremes are generally understood as extremely "positive" or extremely "negative" sentiments, and at least in principle, people are able to order the sentiments of a particular class according to their "positivity." Webster's Dictionary describes the meaning of *positive* in this context as "marked by acceptance or approval" and "indicating agreement or affirmation." Thus, one can replace "positive" and "negative" by "ap-

proving" and "disapproving," respectively. All agents are assumed to order the sentiments of any particular class by the same reflexive, transitive, and complete relation "is at least as approving as."

A complex of sentiments consisting of exactly one sentiment from each class is called a sentiment bundle. An agent's emotional reaction pattern is described by the response function f that, for all real stimulus values s , assigns a sentiment bundle $f(s)$ to s .⁴ The response function is assumed to be continuous and monotonically increasing in the following sense: For all $s^1 > s^0$, all components of $f(s^1)$ are at least as approving as the corresponding components of $f(s^0)$, and at least one is more approving. The more valuable the behavior of others is to the reacting agent, the more approving is the sentiment bundle prompted. Then, f^{-1} is also monotonically increasing, since it maps more approving sentiment bundles to higher stimulus values. This means that f^{-1} is an ordinal approval scale. Hence, the stimulus value s prompting the sentiment bundle $f(s)$ measures the approval associated with $f(s)$.

An agent can obtain approval only from those who come to know his behavior and communicate their feelings to him. It is assumed that these requirements are met exactly by those with whom the agent regularly keeps company: his kin, friends, acquaintances, neighbors, etc. This subgroup is called the reference group of the respective agent and is assumed to be of equal size for all agents. As the model is concerned with symmetric behavior only, the amount of approval obtained from a typical member of the reference group can be used as an index of total approval obtained. Thus, $s(b)$ is the total amount of approval received in return for a cooperative contribution b . Furthermore, agents are assumed to know

the sentiments excited by their behavior⁵ so that, for theoretical purposes, they can be treated as if they knew $s(b)$.

All agents are assumed to have identical preferences for the private good, the collective good, and social approval. With respect to approval preferences, several aspects have to be distinguished. First, it seems to be part of our genetical hardwiring that we are interested in being the objects of others' positive emotions. This means a preference for social approval in the sense of inner states. Second, we also have a preference for the way emotions are expressed. We are not indifferent between, for instance, an invitation for dinner and a bodily attack. One can, however, reasonably assume that the expressive activity is the more favorable the more approving the respective sentiment is. This means that the preference for sentiments as combinations of inner emotional states and expressive activities can be *represented* as a preference for approval. Third, often people not only want to be loved and admired but also to be loved and admired more than others.⁶ I take this into account by assuming a preference for comparative approval $s(b) - s(c)$, where $s(c)$ indicates the approval associated with average behavior. For convenience, it is assumed that absolute and comparative approval can be aggregated into the weighted average $(1 - \alpha)s(b) + \alpha[s(b) - s(c)]$. Thus, the value of the relevant approval variable is

$$(3) \quad a = s(b) - \alpha s(c), \quad 0 \leq \alpha \leq 1.$$

Preferences are represented by a utility function

$$(4) \quad u = u_p(p) + u_c(c) + u_a(a)$$

that satisfies the conventional assumptions

⁴To treat the response function as exogenous does not mean that it is determined exclusively biologically. Certainly, there is some cultural and educational influence. Nevertheless, emotions are essentially not subject to rational control but, rather, are autonomously triggered by the hypothalamus.

⁵This may be justified because of sufficient experience or, as Adam Smith (1759 pp. 9–13) argued, because of "sympathy," the ability to enter into the feelings of others if only the stimulating situation is known.

⁶The assumption for instance is employed by Mandeville, Smith, and Homans. Robert H. Frank (1985) gives a detailed argument for status orientation.

of monotonicity and concavity and, for simplicity's sake, implies essentiality of the private good ($u'_p(0) = \infty$), as well as an absolute elasticity of u'_a smaller than one.

II. Individual Behavior and Social Interaction

A. Individual Behavior

The stimulus power $s(b)$ that measures approval was conceptualized as some subjective value of the cooperative contribution as assigned by the approving agent. The subjective value of a unit contribution, w , is called the "approval rate." In the gift case, the gratitude toward a donor apparently does not only depend on the absolute subjective value of the gift but also upon how this value compares with the values of similar gifts. Analogously, I assume that the effective subjective value that prompts approval is a weighted average of the absolute value wb and the comparative value $w(b - c)$. The corresponding weights are $1 - \beta$ and β so that

$$(5) \quad s(b) = w(b - \beta c), \quad 0 \leq \beta \leq 1.$$

Substituting for $s(b)$ and $s(c)$ from (5) into (3), one obtains

$$(6) \quad a = w(b - \sigma c),$$

$$0 \leq \sigma = \alpha + \beta - \alpha\beta \leq 1.$$

The coefficient σ indicates the strength of the negative externality emanating from the average contribution. For simplicity of language, I treat σ as a measure of "status orientation," although this apparently is only partly correct.

A typical agent is confronted with the respective behavior of others, characterized by w and c , which is beyond his influence. He responds to this behavior by some rationally chosen contribution, b , and some emotionally determined approval rate, v , which he applies to other agents' contributions. The agent's optimal contribution max-

imizes

$$(7) \quad u = u_p(\pi - b) + u_c(c) + u_a[w(b - \sigma c)], \quad 0 \leq b \leq \pi$$

for given w and c with respect to b . The following necessary (and sufficient) condition characterizes an optimum:⁷

$$(8) \quad u'_p(\pi - b) \geq wu'_a[w(b - \sigma c)]$$

and equality if $b > 0$.

Exploiting the assumed properties of the utility function, I routinely obtain the following results from (8):

PROPOSITION 1: *Optimization defines an individual contribution function $b(w, \sigma c, \pi)$ with (i) $b > 0$ if and only if $u'_p(\pi) < wu'_a(-w\sigma c)$ and (ii) $b_w, b_{(\sigma c)}, b_\pi > 0$ and $b_{(\sigma c)}, b_\pi < 1$ for all $b > 0$.*

Condition (i) means that the material incentive for cooperation, w , must be sufficiently large in order to induce cooperative behavior. Formally, the restrictions on $b_{(\sigma c)}$ and b_π are due only to the normalcy of both the private good and approval, whereas $b_w > 0$ results from a dominant substitution effect. It is remarkable that individual contributions are related positively to the degree of status orientation as well as to other agents' contributions, provided σ and c are positive. Increased status orientation makes "keeping up with the Joneses" more important, and increased contributions of others induce efforts to regain at least some of the status lost. Positively related individual contributions are also experimentally observed by James H. Bryan and Mary Ann Test (1967).

In order to develop a hypothesis about the individual approval rate v , I start with the following problem: if individual contributions have only negligible effects on the supply of the collective good and, therefore,

⁷Because of $u'_p(0) = \infty$, the case $u'_p \leq wu'_a$ at $b = \pi$ is impossible.

on other agents' well-being, why then should anybody regard these activities as valuable? Actually, we generally approve of cooperative behavior even if it does not make us significantly better off. In doing so, we often seem to consider the hypothetical advantage we would enjoy if everybody else behaved cooperatively in like manner. This motivates the assumption that an agent's approval rate, his subjective value of another agent's marginal contribution stimulating approval, is equal to the hypothetical advantage, measured in terms of the private good, that the former agent would enjoy if not only the latter but also all other agents except him increased their contributions marginally. Formally, this means that v is taken to be the marginal rate of substitution between π and c with respect to (7):

$$(9) \quad v = \frac{u'_c(c) - \sigma w u'_a[w(b - \sigma c)]}{u'_p(\pi - b)}.$$

Thus, the individual approval rate is the aggregate value of two externalities, the positive collective good externality and the negative status externality. This seems to be supported by the casual observation that people sometimes do not unambiguously approve of contributions to the collective good because, in their opinion, the contributors try to distinguish themselves. For a full understanding of the role of status orientation, it should be noted that, on the one hand, a high σ furthers cooperation for a given incentive w (compare Proposition 1), but on the other hand, it adversely affects cooperation by weakening incentives.

B. Social Equilibrium

A BC equilibrium is an average contribution consistent with individual contributions in the sense that $c = b(w, \sigma c, \pi)$. Substituting $b = c$ into (8) and again exploiting the properties of the utility function, one obtains:

PROPOSITION 2: *In BC equilibrium, individual contributions and supply of the collective good are a function $c(w, \sigma, \pi)$ with (i)*

$c > 0$ if and only if $w > u'_p(\pi)/u'_a(0)$ and (ii) $c_w, c_\sigma, c_\pi > 0$ for all $c > 0$.

By and large, $c(w, \sigma, \pi)$ behaves like $b(w, \sigma c, \pi)$. It can be shown that social interaction reinforces the effects of parametric variations.

A VW equilibrium is an approval rate consistent with individual behavior in the sense that $v = w$. A simultaneous BC and VW equilibrium is called a social exchange equilibrium. Substituting from (8) into (9) and observing $b = c$ and $v = w$, one obtains the following VW condition for a social exchange equilibrium:

$$(10) \quad w = \begin{cases} \frac{u'_c(0)}{u'_p(\pi) + \sigma u'_a(0)} & \text{if } c = 0 \\ \frac{u'_c(c)}{u'_p(\pi - c)} - \sigma & \text{if } c > 0. \end{cases}$$

Figure 1 demonstrates the existence of a unique social equilibrium (w^*, c^*) . The BC curve is positively sloped (compare Proposition 2), whereas concavity of the utility function makes the VW curve negatively sloped. If, for all positive levels of cooperation, the approval rate sustainable at some level of symmetric cooperation is below the one required for supporting the respective level of cooperation, that is, if the VW curve is below the BC curve for all $c > 0$, one obtains $c^* = 0$. It is easily shown that in this case, the approval rate resulting from (10) at $c = 0$ supports a BC equilibrium with $c = 0$.

An increase in π shifts the BC curve to the right and the VW curve upward. Not surprisingly, higher endowments are associated with a higher equilibrium level of cooperation. In general at least, the role of status orientation is unclear. An increasing σ also shifts the BC curve to the right except at $c = 0$, but the VW curve is shifted downward by $\Delta\sigma$. The weakening of the incentive mechanism obviously dominates if c^* is sufficiently small so that low equilibrium levels of cooperation tend to be reduced by increased status orientation. The following proposition collects the results.

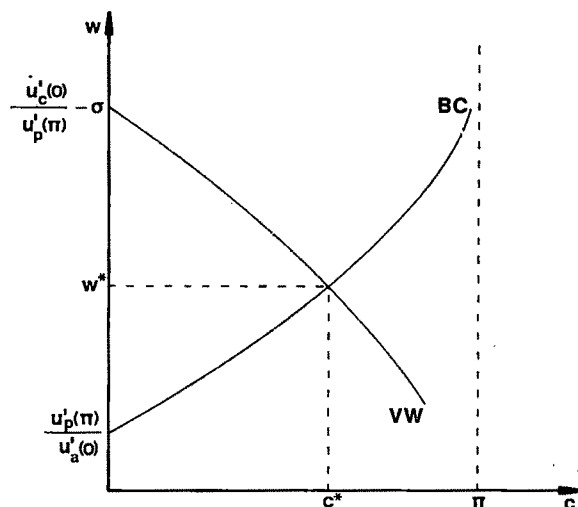


FIGURE 1. SOCIAL EXCHANGE EQUILIBRIUM

PROPOSITION 3: *For given π and σ , there exists a unique social equilibrium (w^*, c^*) with (i) $w^* > 0$, (ii) $c^* > 0$ if and only if*

$$\sigma < \frac{u'_c(0)}{u'_p(\pi)} - \frac{u'_p(\pi)}{u'_a(0)}$$

and (iii) $c_\pi^ > 0$ and, provided c^* is sufficiently small, $c_\sigma^* < 0$.⁸*

III. An Invisible-Hand Explanation of Norm Emergence

In sociological theory, three characteristics of valid behavioral norms seem to be uncontroversial. First, there is a standard of behavior shared by members of a group. The standard is positive in the sense that actual behavior conforms to the standard, at least on average, and normative in the sense that it expresses a shared value judgment as to how group members ought to behave.

⁸In addition, it is easily seen that increasing preferences for approval and the collective good shift the BC curve to the right and the VW curve upward, respectively, whereas increasing preferences for the private good shift the BC curve to the left and the VW curve downward. Thus, if the private good is interpreted as available time and if Staffan B. Linder's (1970) argument of an increasing marginal utility of time due to increasing consumption possibilities is correct, people can be expected to behave more selfishly.

Second, negative or positive individual deviations from the standard are punished or rewarded by negative or positive sanctions, respectively. Third, group members determine their behavior on the basis of the existing standard and anticipated sanctions. However, the problem of sociological theories of norms is the lack of explicit microfoundations. There is no explicit theory of sanction activities or of the relation between sanctions and behavior toward the collective good, not to mention a theory of social interdependence of individual activities. As norms are unintended social results of individual actions, the lack of explicit microfoundations means that sociological theory has little to say as to why and how particular behavioral norms emerge or fall into decay.

The model presented shows how social interaction brings about a standard of behavior c^* to which everybody conforms and how this standard relates to preferences and exogenous variables such as σ and π . Associated with this standard is a "normal" amount of approval $a^* = (1 - \sigma)w^*c^*$. Sanctions are appropriately defined as approval deviations from normal approval that are due to deviations of actual from standard behavior. Thus, deviant behavior $b - c^*$ induces a sanction of amount $w^*(b - c^*)$. The existence of general sanctions also implies

that the behavioral standard reflects the normative expectations of group members with respect to behavior of others. If somebody is sanctioned negatively or positively, this apparently expresses the opinion of others that he deviated negatively or positively from how he ought to behave. It should be clear, then, that the individual optimization problem of the model can be reformulated equivalently in terms of the standard c , deviant behavior $b - c$, normal approval $(1 - \sigma)wc$, and sanctions $w(b - c)$. Thus, the model provides an understanding of norm-regulated behavior.

It is often assumed that norms are internalized. For instance, Talcott Parsons (1951 pp. 38–40) argues that a full institutionalization of a behavioral standard requires its general internalization, by which he means its incorporation into the agents' superego. According to psychoanalytic personality theory, the superego is a major part of the psyche that represents a person's behavioral ideal in agreement with the standards of society and evaluates the correctness of one's behavior from the point of view of this ideal. What is called conscience is held to be manifested by part of the superego. It is supposed to develop in response to advice, warnings, and punishment by parents and other socialization agents. Therefore, internalization of norms essentially means the development of an internal system of sanctions that is structurally equivalent to the external system and, by implication, the adoption of external standards of behavior as own standards. The internal system of sanctions operates in terms of self-approval, which, analogously to external approval, measures the "positivity" of welfare-relevant emotional reactions toward one's own behavior. The possibility of internalization is important to the domain of the model. Insofar as norms are internalized, it may help to understand norm conformity even in situations when the respective agent is unobserved.⁹ A functional equivalent to intro-

jection is the projection of the system of sanctions on an omniscient and almighty god. In the first case, the group manifests itself as part of the individual, in the second as part of God.

IV. Alternative Modes of Allocation and Welfare

A. *Pareto-Suboptimality of the Equilibrium Allocation*

The approval constraint (6) can be interpreted as a household production function describing feasible transformation opportunities of private goods into approval. As the average contribution, c , affects approval generation, all "production functions" are interdependent. First, and not surprisingly, status orientation induces a negative externality. A second externality affects the approval rate through c [compare (9)]. Its sign seems to be unclear. Furthermore, as average contribution and supply of the collective good coincide, there is also a positive externality from production to collective good consumption. In general, these three externalities will render the equilibrium allocation inefficient. In order to establish an efficient symmetric allocation, it is obviously necessary and sufficient to maximize utility as stated in (7) with respect to c and w subject to the symmetry conditions $b = c$ and $v = w$, as well as the approval rate condition (9). In order to realize a first-best allocation systematically, agents would have to be committed to an ethical principle obliging them to adopt the above program. For example, the "Kantian" rule obliging everybody to make the minimal cooperative contribution that he wishes all others to make¹⁰ is obviously sufficient, provided they are well-informed about the structure of the model. However, self-interested rational agents apparently do not follow this rule. They end up in a social exchange equilibrium that, as long as other modes of allocation are ignored, is only second-best.

⁹Robert Trivers (1971 p. 50) argues that humans could have evolved a conscience as a protection against unfavorable consequences of being caught defecting. Frank (1987) analyzes the problem of reliably signaling a conscience.

¹⁰This rule has been called "Kantian" by Jean-Jacques Laffont (1975) and Collard (1978). Others have named it "rational commitment" (John C. Harsanyi, 1980) or "unconditional commitment" (Sugden, 1984).

B. Government and Market Allocation of the Collective Good

Consider now an otherwise identical group without emotional incentives so that approval is zero for everybody. Let the government supply an optimal amount of the collective good financed by enforced symmetric tax payments. The optimal supply of the collective good, c^0 , maximizes utility subject to $b = c$ and zero approval with respect to c . The following necessary and sufficient condition characterizes the optimum:¹¹

$$(11) \quad u'_c(c^0) \leq u'_p(\pi - c^0) \\ \text{and equality if } c^0 > 0.$$

Of course, this is the Samuelson condition. If $c^0 > 0$, the individual willingness to pay, u'_c/u'_p , is equal to 1 so that the aggregate willingness is equal to n , the marginal cost of the collective good. Thus, provided the group had invented some turnstile enabling exclusion from consumption at negligible cost, the above allocation could also be realized as a market equilibrium. The price of one unit of the private good apparently equates supply and demand for the collective good at c^0 .

One might perhaps expect that state or market provides more of the collective good than a system of social exchange. Surprisingly, this is not true a priori. Figure 2 depicts the case $c^* > c^0 > 0$.

As only the BC curve depends on approval preferences, one can apparently conclude that, provided $\sigma < 1$, sufficiently strong desire for approval leads to $c^* > c^0$. (Note that extreme status orientation indeed precludes $c^* > c^0$.) In the light of the model, the empirical finding of Titmuss (1971) that a system of voluntary blood donations provides not less transfusion blood than a blood market is not paradoxical at all.

With respect to group welfare, the optimal allocation attainable through government intervention or an ideal market is not

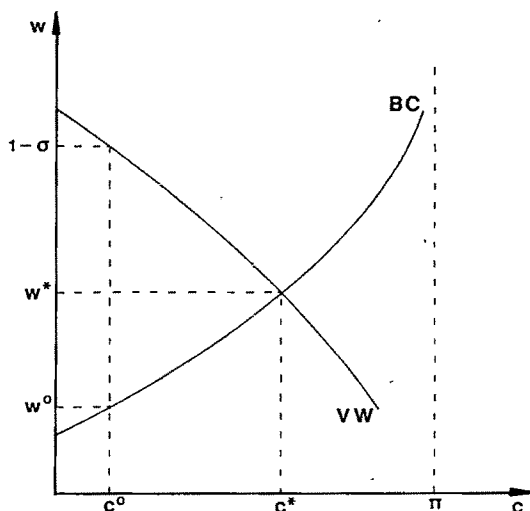


FIGURE 2. ALTERNATIVE ALLOCATIONS

superior to the Kantian allocation and is inferior in general. This is due to the fact that Kantian agents can contribute c^0 voluntarily and thus additionally enjoy nonnegative approval. However, the following result is remarkable.

PROPOSITION 4: *Consider an economy without emotional incentives, but otherwise identical. In the optimal allocation attainable through government intervention or an ideal market, group members are worse off than in the social exchange allocation if the latter provides more of the collective good than the former.*¹²

¹²For proof's sake, consider welfare in BC equilibria. At (w^0, c^0) in Figure 2, social exchange welfare is not less than welfare in a state or market allocation because of nonnegative approval. To complete the argument, I only have to show that the transition from (w^0, c^0) to (w^*, c^*) along the BC curve increases welfare. An increase in w with c remaining constant increases approval ($\sigma < 1$) and, thus, welfare. Now, a marginal increase in the approval rate also increases everybody's contribution and, therefore, c . The increase in contributions is welfare-relevant only through the status and the collective good externality because, owing to optimization, internalized utility effects sum to zero. The aggregate value of a marginal increase in both externalities, measured in terms of the private good, is given by the ordinate value of the VW curve. The latter is positive for all c between c^0 and c^* , which completes the proof.

¹¹Again because of $u'_p(0) = \infty$, the case $u'_c \leq u'_p$ at $c^0 = \pi$ is impossible.

A few remarks are in order. Modes of allocation that rely on voluntary contributions to the collective good have an advantage over other modes. As, from the social point of view, approval is a free by-product of the collective good, the group, compared to other modes, uniformly obtains more outputs from the same inputs, provided the latter are positive and status orientation is not extreme. The advantage in production explains why social exchange, although stuck with externality problems, may do better than an ideal market or a central planner. Even if the supply of the collective good in social equilibrium falls short of c^0 , this may be more than compensated by a positive emotional climate within the group. Whether or not such a compensation actually takes place strongly depends on the degree of status orientation. A group that relies on social exchange for provision of collective goods should try to reduce σ by educating its children 1) not to seek status and prestige and 2) to reward others by positive sentiments for their cooperative contributions, however small they may be, and whatever third agents may contribute.

C. Norms and Expanding Market Domain

Hirsch (1976 Ch. 6) has put forward the interesting hypothesis that empirically observable commercialization of collective goods has the following two related effects. First, market exchange tends to generalize itself in the sense that it considerably weakens or even completely destroys social norms demanding cooperation for the provision of the respective collective goods. Second, this supersession of norms by markets may well decrease social welfare. However, Hirsch would seem to have no explicit model of the interdependence of norms and markets so that it remains unclear whether or not his argument is conclusive. In the following, I show that my model lends support to the commercialization effects as put forward by Hirsch.

Suppose that a new turnstile permits setting up a market for the collective good and that it is supplied infinitely elastically at the marginal-cost price of one unit of the private consumption good. If $c^* \geq c^0$, the

willingness to pay, owing to $u'_c(c^*) \leq u'_p(\pi - c^*)$, is smaller than its marginal-cost price of 1 for any positive market demand, so that market trade is impossible. By contrast, if $c^* < c^0$, every agent will demand $c^0 - c^*$ units of the collective good in the market, because its price is smaller than the initial willingness to pay. Thus, in the case $c^* < c^0$, whatever may be supplied voluntarily of the collective good, its total equilibrium supply must be c^0 if social exchange is combined with market exchange. The same is obviously true if the government intervenes for the collective good. The following result is proved in the Appendix.

PROPOSITION 5: *Consider an economy with $c^0 > c^* > 0$. Then, the opening of a market, as well as government intervention for the collective good, reduces voluntary contributions, social approval if $\sigma < 1$, and possibly also group welfare.*

This can be understood intuitively as follows. In the new equilibrium, a larger part of the private good endowments is allocated to the collective good, so that the marginal utility of private good consumption is increased. This increase has two negative effects. First, it induces people to reduce voluntary contributions for given approval incentives. Second, together with the decrease in the marginal utility of the collective good due to increased supply, it in fact brings down the approval rate, because an increase in collective good consumption is now less valuable in terms of the private good. Both effects cause people to behave more selfishly. Moreover, weakened approval incentives and lower voluntary contributions mean less social approval. Finally, the welfare loss from the colder social climate need not be compensated by the welfare gain from allocating more resources to the collective good.

VI. Concluding Remarks

As the Introduction already provided a summary, I conclude with a few remarks. The analysis presented obviously relies on a number of restrictive assumptions. Some of them have been employed merely to keep

the model as simple as possible. At the expense of rapidly growing complexity, it can be generalized straightforwardly to cover many collective goods and various kinds of asymmetries, for instance. At least two assumptions, namely, the assumed ability to react emotionally to behavior of others and the assumed dependence on emotions and attitudes of others toward oneself, are fundamental and indispensable. Although emotional reactivity and emotional dependence hardly can be doubted empirically, a complete theory of cooperation should also explain how the genes for reactivity and dependence might have spread over the human gene pool. This problem is far from trivial. Emotional reactions will require some energy, and emotional dependence renders the respective agents exploitable by others who are not dependent, so that carriers of either type of gene suffer from an uncompensated disadvantage in reproductive fitness as long as everybody benefits in like manner from the collective good. Thus, in order to explain reactivity and dependence as evolutionary-stable phenomena, one has to give reasons for a comparatively higher collective good consumption of reactive and dependent individuals. Presumably, this will be possible only in a small group setting that actually has been relevant for the evolution of human traits because, compared to the history of evolution, the history of large social formations is negligibly short. Even nowadays, the bulk of voluntary cooperation occurs in small social units such as families, friendships, neighborhoods, and groups of workmates.¹³ Trivers (1971 pp. 48–9) has argued that, within a small group setting, gratitude, moral aggression, and sympathy have been selected for in support of reciprocal altruism, as cooperation is called in biological terminology, but

his reasoning would not seem to be fully convincing. Lack of space makes it impossible to pursue this matter any further here, but work of Axelrod (1984) and Frank (1987) on evolutionary problems structurally closely related to ours suggests that a solution is possible.

APPENDIX Proof of Proposition 5

Let c^0 denote total equilibrium supply of the collective good (as well as total individual contributions), \bar{c} denote voluntary equilibrium supply (as well as voluntary individual contributions), and \bar{w} denote the equilibrium approval rate. The conditions for market equilibrium [compare (11)], approval rate sustainability [compare (10)], and optimization [compare (8)] are

$$(A1) \quad u'_p(\pi - c^0) = u'_c(c^0)$$

$$(A2) \quad \bar{w} = \begin{cases} \frac{u'_c(c^0)}{u'_p(\pi - c^0) + \sigma u'_a(0)} & \text{if } \bar{c} = 0 \\ \frac{u'_c(c^0)}{u'_p(\pi - c^0)} - \sigma & \text{if } \bar{c} > 0 \end{cases}$$

$$(A3) \quad u'_p(\pi - c^3) \geq \bar{w} u'_a[(1 - \sigma)\bar{w}\bar{c}]$$

and equality if $\bar{c} > 0$.

From (10) and (A2), one can conclude $\bar{w} < w^*$ owing to $c^0 > c^*$. Next, I want to show $\bar{c} < c^*$. This is trivial if $\bar{c} = 0$. For $\bar{c} > 0$, (A3) holds with equality. The corresponding condition for c^* is

$$(A4) \quad u'_p(\pi - c^*) = w^* u'_a[(1 - \sigma)w^*c^*].$$

Now, switching from (A4) to (A3), the increase from c^* to c^0 increases the left-hand side, and the decrease from w^* to \bar{w} decreases the right-hand side. In order to compensate for these effects, \bar{c} must be smaller than c^* . Thus, voluntary contributions are lower in the new equilibrium. Owing to $\bar{w} < w^*$ and $\bar{c} < c^*$, social approval is reduced by the amount $(1 - \sigma)(w^*c^* - \bar{w}\bar{c})$. Finally, the welfare loss from reduced approval evidently need not be compensated by the welfare gain from allocating more private resources to the collective good.

REFERENCES

- Arrow, Kenneth J., "Optimal and Voluntary Income Distribution," in S. Rosefielde, ed., *Economic Welfare and the Economics*

¹³ It is quite obvious that the main results of the model are valid also for the small-group case where the affectability of collective good supply only provides an additional incentive for cooperation. Apart from theoretical simplicity, the main reason for choosing a large group setting was to show that even in the most difficult case the approval incentive alone may be sufficient to support significant cooperation.

- of *Soviet Socialism: Essays in Honor of Abram Bergson*, Cambridge: Cambridge University Press, 1981, 267–88.
- Axelrod, Robert, *The Evolution of Cooperation*, New York: Basic Books, 1984.
- Becker, Gary S., "A Theory of Social Interaction," *Journal of Political Economy*, November/December 1974, 82, 1063–93.
- , "Altruism, Egoism, and Genetic Fitness," *Journal of Economic Literature*, September 1976, 14, 817–26.
- Bryan, James H. and Test, Mary Ann, "Models and Helping: Naturalistic Studies in Aiding Behavior," *Journal of Personality and Social Psychology*, 1967, 8, 400–7.
- Collard, David A., *Altruism and Economy*, Oxford: Martin Robertson, 1978.
- Frank, Robert H., *Choosing the Right Pond*, New York: Oxford University Press, 1985.
- , "If *Homo economicus* Could Choose His Own Utility Function, Would He Choose One with a Conscience?" *American Economic Review*, September 1987, 77, 593–604.
- Hammond, Peter, "Charity: Altruism or Cooperative Egoism?" in E. Phelps, ed., *Altruism, Morality, and Economic Theory*. New York: Russell Sage Foundation, 1975, 115–31.
- Harsanyi, John C., "Rule Utilitarianism, Rights, Obligations and the Theory of Rational Behavior," *Theory and Decision*, March 1980, 12, 115–33.
- Hirsch, Fred, *Social Limits to Growth*, Cambridge, MA: Harvard University Press, 1976.
- Homans, George C., *Social Behavior. Its Elementary Forms*. New York: Harcourt Brace Jovanovich, 1961.
- Kitcher, Philip, *Vaulting Ambition: Sociobiology and the Quest for Human Nature*, Cambridge, MA: MIT Press, 1985.
- Kurz, Mordecai, "Altruistic Equilibrium," in B. Belassa and R. Nelson, eds., *Economic Progress, Private Values, and Public Policy*, Amsterdam: North Holland, 1977, 177–200.
- Laffont, Jean-Jacques, "Macroeconomic Constraints, Economic Efficiency and Ethics: An Introduction to Kantian Economics," *Economica*, November 1975, 42, 430–47.
- Linder, Staffan B., *The Harried Leisure Class*. New York: Columbia University Press, 1970.
- Mandeville, Bernard de, *The Fable of the Bees*, F. B. Kay, ed., Oxford: Oxford University Press, 1924, originally published 1714.
- Margolis, Howard, *Selfishness, Altruism, and Rationality. A Theory of Social Choice*, Cambridge: Cambridge University Press, 1982.
- Mueller, Dennis C., "Rational Egoism Versus Adaptive Egoism as Fundamental Postulate for a Descriptive Theory of Human Behavior," *Public Choice*, 1986, 51, 3–23.
- Olson, Mancur, *The Logic of Collective Action. Public Goods and the Theory of Groups*, Cambridge, MA: Harvard University Press, 1965.
- Parsons, Talcott, *The Social System*, London: Routledge and Kegan Paul, 1951.
- Schotter, Andrew, *The Economic Theory of Social Institutions*. Cambridge: Cambridge University Press, 1981.
- Schwartz, Robert, "Personal Philanthropic Contributions," *Journal of Political Economy*, November/December 1970, 78, 1264–91.
- Smith, Adam, *The Theory of Moral Sentiments*, in D. D. Raphael and A. L. Macfie, eds. Oxford: Clarendon Press, 1976, originally published 1759.
- Sugden, Robert, "On the Economics of Philanthropy," *Economic Journal*, June 1982, 92, 341–50.
- , "Reciprocity: The Supply of Public Goods Through Voluntary Contributions," *Economic Journal*, December 1984, 94, 772–87.
- Titmuss, Richard, *The Gift Relationship*, London: Allen & Unwin, 1971.
- Trivers, Robert, "The Evolution of Reciprocal Altruism," *Quarterly Review of Biology*, March 1971, 46, 32–57.
- Wilson, Edward O., *Sociobiology*, Cambridge, MA: Harvard University Press, 1975.

The Economic Effects of Production Taxes in a Stochastic Growth Model

By MICHAEL DOTSEY*

This paper analyzes the effects of stochastic taxes on production in a simple stochastic growth model. In so doing, the paper explicitly examines the importance of uncertainty on the decisions of individual agents. This importance is emphasized by comparing economic outcomes to various realizations of tax rates under uncertainty and perfect foresight. (JEL 321)

The Economic Recovery Tax Act of 1981 led to the largest postwar decline in effective tax rates on capital. The legislation also had its most significant effect on rates in 1982 because of the rapid decline in inflation. Although some of the tax cut was rescinded in 1982, effective corporate tax rates on plant and equipment, measured as the difference between before- and after-tax rates of return to capital as a percentage of before-tax rates of return, remained at historically low values through 1986. Accompanying this tax cut is the current economic recovery, which began in November 1982. During this recovery we have witnessed relatively large increases in business fixed investment, a stock market boom, and a large rise in both the *ex post* and *ex ante* real interest rate. It is, therefore, natural to investigate the linkages between the tax cut and the increase in economic activity.

The effects of this particular tax cut have been consistent with a pattern of negative correlations between taxes and real interest rates, stock prices, investment, and output growth observed following previous business tax cuts. For example, using annual data on

the effective tax rates on plant and equipment reported in Charles R. Hulten and James W. Robertson (1982), the correlation coefficients between the logarithm of tax rates and the logarithms of real GNP, real business fixed investment, and the New York Stock Exchange price index for the period 1952–84 are -0.56 , -0.55 , and -0.65 . Further, using an autoregression to calculate expected inflation over the period 1960–84, the logarithm of one plus the *ex ante* after-tax real rate of interest and the logarithm of the effective tax rate display a correlation coefficient of -0.38 , while the coefficient with respect to the logarithm of one plus the *ex post* real rate is -0.50 .¹

While the general effects of the recent tax cut are consistent with effects observed in other periods, the relative size of the 1981 tax cut is quite large. For instance, Hulten and Robertson calculate that the effective tax rate on capital in nonresidential business was reduced from roughly 33 percent in 1980 to approximately 1 percent in 1984. It is not surprising that a change of this magnitude has generated renewed interest in the interaction between taxes on capital and real economic variables.

This paper examines the substitution effects of tax rate changes within the context of a stochastic growth model. Because taxes drive a wedge between the marginal product of capital and its after-tax return the

*Federal Reserve Bank of Richmond, 701 E. Byrd St., Richmond, VA 23219. I am indebted to John Boyd, James Hamilton, Tony Kuprianov, Ching Sheng Mao, Paul Romer, Alan Stockman, and two anonymous referees for valuable comments and discussions during the course of this research. Bob LaRoche provided excellent research assistance. The views expressed in this paper are solely those of the author and do not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

¹When first differences of the logarithm of real GNP and real business fixed investment are used, the correlation coefficients are -0.38 and -0.06 .

resulting equilibrium is suboptimal. Further, the effects of taxes are highlighted by the lump-sum remittance of tax proceeds to individuals. The methodology for finding the solutions to an individual's dynamic programming problem relies on the envelope theorem and the construction of a policy function that simultaneously solves the individual's optimization problem and market clearing conditions. Attacking the problem in this way avoids explicit consideration of the value function and produces a tractable method of analysis for problems in which aggregate conditions affect an individual's value function. Essentially, the problem is reduced to solving one functional equation for policy.

With the exception of David Bizer and Kenneth L. Judd (1988), most of the work dealing with the effects of taxes has proceeded within the confines of standard non-stochastic growth models (e.g., William A. Brock and Stephen J. Turnovsky, 1981; Andrew B. Abel and Oliver Blanchard, 1982) or in models in which agents have perfect foresight regarding the path of tax rates (e.g., Robert A. Becker, 1985; Lawrence H. Goulder and Lawrence H. Summers, 1987). Little effort seems to have been given to examining the effects of tax rate changes when taxes are explicitly depicted as following a particular stochastic process.

This paper takes the latter approach and investigates the effects of tax changes in a stochastic growth model in which only tastes, technology, and the stochastic process for taxes are exogenous. This procedure takes seriously the methodology advocated by Robert E. Lucas, Jr. (1976) and Thomas J. Sargent (1979) that a policy should be represented as a given outcome of some stochastic process. The analysis yields investment, output, and real interest rate behavior that depend explicitly on tastes and technology parameters as well as the underlying process generating tax rates.

The qualitative movements in endogenous variables that are generated by tax rate changes in this model are similar in some instances to results derived in the nonstochastic or perfect foresight models. For instance, when a high tax rate is gener-

ated (and the rate is expected to remain high), agents reduce the capital stock. This leads to lower real interest rates and lower transitional real output growth. The qualitative similarity of the results from the various models is a natural outcome of the behavior of agents, behavior that is characterized by analogous intertemporal optimization problems in all three classes of models.

An explicit stochastic treatment of taxes allows one to examine the effects of uncertainty on the decisions of individual agents. This uncertainty should be an important consideration in determining behavior, since tax law changes are fairly frequent and changes in inflation do result in significant movements in the effective tax rate on capital. Also, the exact nature of the uncertainty is related to the specific process that taxes are assumed to follow. For example, the degree of persistence of the process generating tax rates will have important implications for individual behavior. Therefore, agents will show quantitatively different behavior for any specific realizations of tax rates when the inherent randomness of taxes is modeled as opposed to treating tax rate data as being known with certainty. If one wants to derive realistic decision rules, then the stochastic nature of the agents problem needs to be analyzed explicitly.

The results generated by the stochastic growth model derived in this paper produce correlations that are consistent with those mentioned above. However, the simple model examined below is not overly successful at replicating correlation coefficients at various leads and lags, especially when the first differences of logs are used. A more detailed investigation that tests to see if incorporating tax policy into a more sophisticated model improves that models ability to replicate actual time series is needed before one can determine the importance of taxes on economic activity.

The paper is structured as follows. Section I contains a general description of the model, while Section II examines a particular solution for the case of log utility and exponential production. Of particular interest is the derivation of a closed-form solution to the nonlinear stochastic difference

equations that determine the equilibrium. In this section, the importance of the stochastic process generating taxes is emphasized by examining the cases where taxes follow either a two-state Markov process or are independently and identically distributed. In Section III, the model is extended to examine investment tax credits. Since doing so does not substantially affect the results, the treatment is fairly concise. A comparison between capital stocks generated by the model for particular realizations of a stochastic process under uncertainty with those generated by a perfect foresight model is made in Section IV, while a short summary is given in Section V.

I. The Model

The model is a one-sector stochastic growth model consisting of three economic entities: firms, consumers, and the government. Individuals are infinitely lived and maximize the discounted stream of momentary utility while firms produce output according to a concave production function, $f(k)$, where output is total output that is available for production and consumption.² The various economic entities interact in two markets each period. First, there is a capital market in which firms purchase capital from individuals. Capital is carried over from the previous period and is therefore supplied inelastically as in Brock (1979). Next there is a combined goods and securities market in which individuals allocate their wealth among goods and securities. Individuals also decide how much they will consume and how much capital to carry into the succeeding period. The government taxes away some of the firm's revenue and remits the proceeds lump sum to individuals. Tax rates are stochastic and are announced at the beginning of each period so that there is no uncertainty over the current tax rate, but there is uncertainty over future tax rates.

²One could easily think of $f(k)$ as equaling total production $g(k)$ minus depreciation $D(k)$.

A. Capital Market

At the beginning of period t , the representative individual has carried over k_t units of capital that are sold to the firm for p_t units of output per unit of capital. One can think of output as seeds that can either be eaten (consumed) or invested (carried into the next period and sold to firms). In this respect the model is similar to Becker (1985) and Jean Pierre Danthine and John B. Donaldson (1985). The firm maximizes its after-tax profits

$$(1) \quad d_t = (1 - \tau_t)f(k_t) - p_t k_t$$

where τ is the effective tax rate on capital. The effective tax rate on capital is basically an index number that aggregates the effects of various components of the tax code (for example, the legislated tax rate, depreciation allowances, and the effects of inflation). It is essentially an attempt to calculate a number that is sufficient for determining individual behavior in response to various changes in taxation (for a detailed description of its calculation, see Hulten and Robertson [1982]).³

This optimization implies that the price of capital is equated to its after-tax marginal product; therefore, the price of capital and profits are determined as functions of the capital stock and the tax rate. Formally, $p_t = (1 - \tau_t)f'(k_t)$. This formulation of tax on output is chosen for analytical tractability and captures the tax wedge resulting from a given effective tax rate on capital.

B. Goods and Securities Market

After selling capital to firms and receiving the distribution of profits and lump-sum tax remissions, individuals choose their current

³As formulated, equation (1) makes τ appear as a production tax rather than an income tax. As noted in the text, the tax rate that is being used is an effective tax rate, not a legislated rate. Also, even if τ were just the legislated profits tax, since legislated depreciation allowances are not based on true economic depreciation, the current tax code involves aspects of a wealth tax.

consumption c_t , next period's holdings of capital, k_{t+1} , and their share of the firm s_{t+1} subject to the value of their current wealth, w_t , where

$$(2) \quad w_t = (q_t + d_t)s_t + p_t k_t + T_t.$$

The first term after the equality represents the current value of the shares of the firms, $q_t s_t$, plus dividend payments $d_t s_t$. The second term is the payment for capital, and the third term is the per capita lump-sum transfer of tax proceeds. Since all firms have access to the same strictly concave production technology, $f(k)$, each firm will utilize the same amount of capital, \hat{k} . In general for m firms and n individuals (where m and n are large), $T_t = (m/n)\tau_t f(\hat{k}_t)$. For simplicity of exposition, let $m = n$ and hence $T_t = \tau_t f(K_t)$, where K_t is the aggregate per capita capital stock. The budget constraint facing the individual is

$$(3) \quad c_t + k_{t+1} + q_t s_{t+1} \leq w_t.$$

C. The Individual's Maximization

The individual's problem is to maximize his discounted expected utility

$$(4) \quad U = E_t \sum_{j=t}^{\infty} \beta^{j-t} u(c_j)$$

subject to his budget constraint (3).

The solution to the competitive equilibrium for the economy described above is equivalent to the solution of a planner's problem in which the planner faces a stochastic discount rate. In particular, the competitive equilibrium is equivalent to the problem where the planner maximizes

$$(4') \quad U = E_t \sum_{j=t}^{\infty} b^{j-t} u(c_j)$$

subject to a sequence of constraints

$$(3') \quad c_t + k_{t+1} \leq f(k_t)$$

where $b^m = \beta^{m-1} \prod_{k=t}^m (1 - \tau_k)$. This planning problem is the stochastic analogue of the one presented in Becker (1985). The solution involves the maximization of a concave function over a convex set and is, therefore, unique. Generally, it is difficult to find an equivalence between a planner's problem and a competitive equilibrium involving a tax structure that is more involved than the one analyzed here. For that reason, the paper proceeds by examining the solution to the competitive equilibrium directly. The problem is first posed in terms of dynamic programming and is formally stated as

$$(5) \quad V(k, K, s, \tau) \\ = \max\{u(c) + \beta EV(k', K', s', \tau')\} \\ c, k', s'$$

such that

$$c + k' + qs' \leq w$$

$$\text{and} \quad w = (q + d)s + pk + T$$

where per capita capital is determined by $K' = \Psi(K, \tau)$. The price of capital, $p(K, \tau)$, and profits, $d(K, \tau)$, are given by the solution to the firm's problem. Tax rebates are given by $T_t = \tau_t f(K_t)$, and asset prices are determined by the function $q(K, \tau)$. In solving their optimization problem, individuals take as given $q(K, \tau)$, $d(K, \tau)$, and $p(K, \tau)$, as well as the policy function that describes the transition path of the aggregate per capita capital stock $K' = \Psi(K, \tau)$. Further, the notation E is used to denote the conditional-expectations operator where expectations are conditioned on all current information, including the realization of this period's tax rate.

In order to guarantee a unique solution to the value function, one must place fairly severe restrictions on taste and technology. Wilbur John Coleman II (1989) presents a detailed analysis of existence and uniqueness problems in a setting similar to the one

examined here.⁴ These restrictions are not met by the examples considered in Section II. However, one can verify that the solutions presented below satisfy both the Euler equations and the transversality conditions of the planner's problem and are, therefore, the correct solutions.

The first-order conditions for the consumer's problem are given by

$$(6a) \quad Du(c) = \lambda$$

$$(6b) \quad \beta ED_1 V(k', K', s', \tau') = \lambda$$

$$(6c) \quad \beta ED_3 V(k', K', s', \tau') = \lambda q$$

and the budget constraint (3). The notation D_i represents the partial-derivative operator with respect to a function's i th argument, and λ is the Lagrange multiplier associated with the individual's budget constraint.

Equation (6a) implies that the marginal utility of consumption equals the marginal utility of wealth. That is, individuals are indifferent between consuming or holding an extra unit of wealth. Equation (6b) states that the discounted value of next period's marginal utility of wealth times the after-tax marginal productivity of capital equals the marginal utility of wealth. This means that at an optimum the individual is indifferent between investing in an extra unit of capital and consuming less today. Equation (6c) is the difference equation determining the price of equity. It implies an indifference at the margin of selling equity today and holding the equity and selling it next period.

D. Equilibrium

The solution to the competitive equilibrium involves finding the individual's policy function $k' = \psi(k, K, s, \tau)$ that simultane-

ously satisfies his first-order conditions, is consistent with market clearing in the goods and asset markets, and yields individual holdings of capital equal to the aggregate per capita capital stock. Formally, the equilibrium conditions are

$$(7a) \quad s = 1$$

$$(7b) \quad c + k' = f(k)$$

$$(7c) \quad k = K.$$

Because individuals are too small to affect the aggregate capital stock and taxes are distortionary, the solution to the competitive equilibrium is suboptimal.

Solving for the competitive equilibrium involves making use of the envelope theorem to obtain $D_1 V(k, K, s, \tau) = \lambda p$ and $D_3 V(k, K, s, \tau) = \lambda(q + d)$. Equation (6b) can be rewritten as

$$(8a) \quad \beta EDu(c')p' = Du(c).$$

Employing the equilibrium conditions (7a)–(7c) and $\psi(k, k, 1, \tau) = \Psi(k, \tau) = h(k, \tau)$, the solution to the problem involves finding a function h that satisfies

$$(8b) \quad \beta EDu[f(h(k, \tau)) - h(h(k, \tau), \tau')] \\ \times (1 - \tau')f'(h(k, \tau)) \\ = Du[f(k) - h(k, \tau)].$$

Finally, the solution for equity prices is given by

$$(8c) \quad \beta E[\lambda'(q' + d')] = \lambda q.$$

In general, proving the existence of the policy function $h(k, \tau)$ requires the same conditions as those already needed to guarantee the existence of the individual's value function, while uniqueness requires a few additional restrictions (see Coleman, 1989). Since the problem under consideration is equivalent to a planning problem, one need only check that a policy function satisfying (8b) also solves the planner's problem. Upon finding $h(k, \tau)$, it is then possible to con-

⁴The necessary conditions for the existence of a unique V satisfying (5) are (a) momentary utility must be a strictly concave, twice continuously differentiable function that is bounded from below with the property that $u'(0) = \infty$; (b) the production function is also strictly concave and twice continuously differentiable with $f(0) = 0$, $1/\beta < b \leq f'(0) \leq B < \infty$, and there exists a finite \bar{k} such that $f(\bar{k}) \leq \bar{k}$; and (c) finally, $(1 - \tau)f'(k)$ must be strictly decreasing in k .

struct the equilibrium stochastic process for capital, consumption, the price of capital, and security prices.

II. An Example

A. A Particular Solution

Since the main concern of the analysis is to examine how various realizations of tax rates affect economic activity, the remainder of the paper will examine a particular functional form for which a closed-form solution exists. In particular, the utility function $u(c) = \ln(c)$ and the production function $f(k) = k^\alpha$, $0 < \alpha < 1$, will be emphasized.⁵ The results of this section will concentrate on the effects of production taxes, but a brief investigation of investment tax credits will be undertaken in Section III. The closed-form solutions to these problems are nontrivial and involve the solution to a set of nonlinear difference equations.

Without the remittance of taxes (i.e., if tax proceeds were destroyed), this problem would be equivalent to the one solved by Brock (1979), where the production function was subject to multiplicative productivity shocks. In that case, the fraction of output allocated to investment would equal $\alpha\beta$ and be independent of the stochastic process followed by taxes. The assumption that all tax proceeds are remitted lump sum is, to some extent, like assuming that government spending is valued exactly like consumption. This allows one to ignore the effects of government spending and to isolate the effect of taxes. Also, the remittance of these proceeds implies that it is the compensated effects of consumption that are being analyzed and that even in the pres-

ence of log utility, expectations of future taxes will be an important determinant of current decisions.

The procedure used to calculate the competitive equilibrium is to find a policy function that satisfies the particular form of the difference equation given in (8b) that is generated in this example. Alternatively, one could try to find the value function, but this latter procedure does not appear promising. In problems involving suboptimal equilibria where aggregate's state variables appear in the individual's value function, finding a policy function seems to be a precondition to calculating the value function.

With taxes either following a first-order Markov process or distributed independently and identically, an intuitive guess regarding the decision rules governing consumption and investment is that each is a fraction of output and that these fractions are potentially functions of the current realization of the tax rate (past realizations would be important for Markov processes of higher order). That decision rules follow a linear process can be directly derived using the symmetry equilibria procedures in John H. Boyd III (1986).⁶ In particular, $k' = \gamma(\tau)f(k)$ and $c = (1 - \gamma(\tau))f(k)$ where

$$(9) \quad \gamma(\tau) = \frac{\alpha\beta g(1-\tau)}{1 + \alpha\beta g(1-\tau)}$$

and $g(1-\tau)$ is given by the recursive relationship

$$(10) \quad g(1-\tau) = E[(1-\tau')[1 + \alpha\beta g(1-\tau')]].$$

The function $g(1-\tau)$ is unique since it is

⁵Because results using log utility sometimes represent a special case, simulations were run using constant-relative-risk-aversion utility functions with risk-aversion parameters of 1/2 and 2. Also, a production technology incorporating less than full depreciation $f(k) = k^\alpha + (1-\xi)k$ was analyzed. None of the results involving quantities was qualitatively affected either with respect to correlation coefficients or the comparison of capital paths under uncertainty and perfect foresight. I am indebted to Ching Sheng Mao for running simulations using his real-business-cycle model.

⁶The symmetry can be written as $S(p_t, q_t, c_t, s_t, k_t, \pi_t, T_t) = (\lambda_{t+1}\lambda_t^{-1}p_t, \lambda_{t+1}q_t, \lambda_{t+1}c_t, s_t, \lambda_t k_t, \lambda_t \pi_t, \lambda_{t+1}T_t)$, where $\lambda_{t+1} = \lambda_t^{\alpha'}$. This symmetry maps initial capital k into λk , preserves household budget constraints, the firm's objective, the government's budget, the definition of profits, and market clearing. It also implies that the choice of capital and consumption is linear in income and depends on the tax process. I am indebted to John Boyd for describing his methodology to me.



the fixed point of a contraction operator implicitly defined by (10).⁷ That this is a solution to economy-wide equilibrium is shown in the Appendix.

The closed-form solutions for c and k' indicate that the mixture between consumption and investment is based on the conditional expectation of the entire path of future taxes. The expected path of future taxes is relevant since it affects the value of future capital and the amount of consumption that can be purchased from the sale of capital to firms.

B. The Solution When Taxes Are i.i.d. (Independently and Identically Distributed)

Further intuition regarding the economic effects of tax rate changes can be gained by looking at the results obtained when taxes follow a particular stochastic process. To highlight the difference between permanent and transitory tax rate movements, both an i.i.d. and a simple two-state first-order Markov process will be used. The behaviors of investment, equity prices, and the real rate of interest differ quite markedly under the two assumed distributions. These examples, therefore, clearly illustrate the importance of correctly specifying the stochastic process for taxes if one is to have confidence in the derived consequences of tax rate changes.

When taxes are independently and identically distributed, with mean $\bar{\tau}$, certainty equivalence obtains, and $g(1 - \tau_t) = (1 - \bar{\tau}) / (1 - \alpha\beta(1 - \bar{\tau}))$, implying that $\gamma(1 - \tau_t) = \alpha\beta(1 - \bar{\tau})$. Hence the fraction of output devoted to investment is independent of the current realization of taxes, but as in the deterministic tax literature, investment will be lower for higher average tax rates. Further, the time path of consumption and the capital stock are independent of tax realizations as long as tax proceeds are rebated.

Regarding security prices, equation (8c) represents a first-order difference equation

that along with (1), the fact that $u'(c) = \lambda$, and the solutions for $\gamma(1 - \tau)$ and $g(1 - \tau)$ yields⁸

$$(11) \quad q_t = \beta c_t \sum_{j=0}^{\infty} \beta^j E_t \frac{d_{t+j+1}}{c_{t+j+1}} \\ = \beta(1 - \alpha) c_t E_t \sum_{j=0}^{\infty} \beta^j g(1 - \tau_{t+j}).$$

Security prices are observed to equal the stream of after-tax profits discounted by the risk-adjusted real rate of interest. As in Brock (1979), the asset price represents a return to the technology k^α and is directly related to profits share of output $(1 - \alpha)$, as well as to the discount rate β .

Using the solution for $g(1 - \tau)$ implies that

$$(12) \quad q_t = \frac{\beta(1 - \alpha)(1 - \bar{\tau})}{(1 - \beta)} y_t$$

and that equity prices are invariant to tax rate realizations. Similarly, it can be shown that the after-tax real rate of interest is invariant to realizations of the tax rate.

C. The Solution When Tax Rates Are Persistent

When tax rates are no longer independently and identically distributed, the property of certainty equivalence no longer holds. However, with taxes following a Markov

⁸The second equality in (11) is derived using (1), (6b), and the envelope theorem. Using the equation for profits (1) and the consumption function,

$$d_{t+1} / c_{t+1} = (1 - \alpha) \frac{1 - \tau_{t+1}}{1 - \gamma(\tau_{t+1})}.$$

Equation (8b) leads to the difference equation

$$\alpha\beta E_t \frac{1 - \tau_{t+1}}{1 - \gamma(\tau_{t+1})} = \frac{\gamma(\tau_t)}{1 - \gamma(\tau_t)}.$$

From the definition of $g(1 - \tau_t)$, the latter expression is also equal to $\alpha\beta g(1 - \tau_t)$.

⁷This last statement was pointed out by an anonymous referee.

process, the recursive nature of $g(1-\tau)$ can be used to derive closed-form solutions. For example, consider a first-order Markov process with transition probabilities given by

$$(13a) \quad \text{prob}(\tau_{t+1} = \tau_0 | \tau_t = \tau_0) = \pi_0$$

$$(13b) \quad \text{prob}(\tau_{t+1} = \tau_1 | \tau_t = \tau_1) = \pi_1$$

where $0 < \tau_0 < \tau_1 < 1$. Given the discussions in Robert J. Barro (1979) and Robert E. Lucas and Nancy L. Stokey (1983) concerning the use of government debt to smooth tax distortions over time, one would expect the tax rate to show a good deal of persistence, with both π_0 and π_1 being greater than $1/2$ and perhaps close to one. Further, the qualitative results generated by using the simple process described in (13a) and (13b) are not altered by using a Markov process having more than two states. Therefore, the qualitative results yielded by this example are of general interest. Taking advantage of the recursive nature of $g(1-\tau_t)$ yields

$$(14) \quad g(1-\tau_j) = [(1-\tau_j)\pi_j + (1-\tau_i)(1-\pi_j) - \alpha\beta(1-\tau_i)(1-\tau_j)\delta] / \Delta$$

where

$$\Delta = 1 - \alpha\beta[(1-\tau_j)\pi_j + (1-\tau_i)\pi_i] + \alpha^2\beta^2[(1-\tau_i)(1-\tau_j)\delta]$$

for $i, j = 0, 1$, $i \neq j$, and $\delta = \pi_i + \pi_j - 1$.

1. *Consumption and Investment.* The behavior of consumption and investment will now be strikingly different than in the i.i.d. case. For the Markov process under consideration,

$$(15) \quad g(1-\tau_0) - g(1-\tau_1) = \frac{1}{\Delta} [(\tau_1 - \tau_0)(\pi_0 + \pi_1 - 1)].$$

Equation (15) implies that $g(1-\tau_0) > g(1-\tau_1)$ if $\pi_0 + \pi_1 > 1$, that is, if tax rates are likely to persist over time. From the definition of $\gamma(\tau)$, this means $\gamma(\tau_0) > \gamma(\tau_1)$. Hence, a greater fraction of output will be invested in the low tax state if the low tax state implies that taxes are more likely to be low in the future. Also, the difference between $\gamma(\tau_1)$ and $\gamma(\tau_0)$ will be directly related to the degree of persistence, a result that is analogous to behavior reported in Bizer and Judd (1988).

A comparison between these two processes, a first-order Markov process with $\pi_0 + \pi_1 > 1$ and an i.i.d. process, points out the problems that arise if one simply assumes that agents have perfect foresight. The behavior of investment for some given realization of taxes depends crucially on the distribution from which tax realizations were drawn. As will be shown in more detail in Section IV, arbitrarily forcing expectations to equal actual realizations may be misleading and certainly affects the quantitative results of the analysis.

Using the simple nature of the closed-form solutions, the correlation coefficients between tax rates and output, consumption, or the capital stock can be calculated. These correlations may be of more interest than the simple comparative static exercise just conducted, since they facilitate a comparison of the model with actual data. With 100-percent depreciation, the correlation coefficients between taxes and output and between taxes and capital are equivalent, and only the latter is presented. Numerical results for $\alpha = 1/3$, $\beta = 0.97$, $\pi_0 = \pi_1 = 0.9$, $\tau_0 = 1/4$, and $\tau_1 = 1/2$ are presented in Table 1.

Although straightforward, the calculations are cumbersome. To avoid burdening the reader, only the derivation for the covariance between $\hat{\tau}_t = \log(\tau_t)$ and $\hat{k}_{t+1} = \log(k_{t+1})$ is presented. The calculation of covariances at other leads and lags is similar. From the definition of covariance,

$$(16) \quad \text{Cov}(\hat{\tau}_t, \hat{k}_{t+1}) = E(\hat{\tau}_t \hat{k}_{t+1}) - E(\hat{\tau}_t)E(\hat{k}_{t+1}).$$

TABLE 1—CORRELATION COEFFICIENTS BETWEEN $\ln(\tau)$ AND $\Delta \ln y, \Delta \ln(1 + \rho), \Delta \ln q$

	(a) Predicted by the Model ¹			(b) Actual Data ²		
	$x_t = \Delta \ln(y_t)$	$x_t = \Delta \ln(1 + p_t)$	$x_t = \Delta \ln q_t$	$x_t = \Delta \ln(y_t)$	$x_t = \Delta \ln(1 + p_t)$	$x_t = \Delta \ln(q_t)$
$\text{cor}(\ln \tau, x_{t+3})$	0.31	0.016	0.27	0.14	0.17	-0.26
$\text{cor}(\ln \tau, x_{t+2})$	0.17	0.12	0.10	-0.009	-0.03	-0.18
$\text{cor}(\ln \tau, x_{t+1})$	-0.43	0.47	-0.50	-0.22	-0.25	-0.007
$\text{cor}(\ln \tau, x_t)$	-0.35	-0.096	-0.28	-0.38	-0.56	0.10
$\text{cor}(\ln \tau, x_{t-1})$	-0.27	-0.075	-0.23	-0.001	-0.22	0.02
$\text{cor}(\ln \tau, x_{t-2})$	-0.22	-0.066	-0.15	0.19	-0.16	-0.01
$\text{cor}(\ln \tau, x_{t-3})$	-0.18	-0.025	-0.14	0.25	-0.13	-0.10

¹The correlation coefficients from the model are calculated by using parameter values of $\beta = 0.97$, $\alpha = 1/3$, $\Pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}$, $\tau_0 = 0.25$, and $\tau_1 = 0.50$.

²The effective tax rate series is from Hulten and Robertson (1982) for the years 1952–84. The data on *ex ante* expected real interest rates subtracts estimates of expected inflation calculated by using an autoregression from the one-year municipal bond rates. The series extend from 1960 to 1986.

Using the recursive relationship $\hat{k}_{t+1} = \hat{\gamma}(\tau_t) + \alpha \hat{k}_t$ and the stationarity of the model, (16) can be written as

$$(17) \quad \text{Cov}(\hat{\tau}_t, \hat{k}_{t+1}) \\ = E[\hat{\tau}_t(\hat{\gamma}(\tau_t) + \alpha \hat{\gamma}(\tau_{t-1}) \\ + \alpha^2 \hat{\gamma}(\tau_{t-2}) + \dots)] \\ = \frac{E(\hat{\tau})E(\hat{\gamma}(\tau))}{1 - \alpha}$$

Using the Markov property of τ and γ yields

$$(18) \quad \text{Cov}(\hat{\tau}_t, \hat{k}_{t+1}) \\ = \left(\frac{1}{2} \right) (\hat{\tau}_0 \hat{\tau}_1) (\mathbf{I} + \alpha \mathbf{\Pi} + \alpha^2 \mathbf{\Pi}^2 + \dots) \begin{pmatrix} \hat{\gamma}(\hat{\tau}_0) \\ \hat{\gamma}(\hat{\tau}_1) \end{pmatrix} \\ = \frac{E(\hat{\tau})E(\hat{\gamma}(\tau))}{1 - \alpha}$$

where $\mathbf{\Pi}$ is the probability transition matrix $\begin{pmatrix} 1-\pi & \pi \\ \pi & 1-\pi \end{pmatrix}$. The solution reduces to

$$(19) \quad \text{Cov}(\hat{\tau}_t, \hat{k}_{t+1}) \\ = \frac{1}{4} \frac{(\hat{\tau}_0 - \hat{\tau}_1)(\hat{\gamma}(\tau_0) - \hat{\gamma}(\tau_1))}{1 - \alpha \delta}$$

where $\delta = 2\pi - 1$. Calculating the variance of $\hat{\tau}$ is straightforward, while deriving the variance of \hat{k}_{t+1} is somewhat more involved, since it equals $[1/(1 - \alpha^2)](\text{Var}(\hat{\gamma}(\tau)) + 2\text{Cov}(\hat{\gamma}(\tau), \hat{k}_{t+1}))$. Letting $r(j) = \text{correlation}(\hat{\tau}_t, \hat{k}_{t+j})$ and using the above parameter values, one finds that $[r(3), r(2), r(1), r(0), r(-1), r(-2), r(-3)]$ equals $[-0.76, -0.90, -0.98, -0.78, -0.62, -0.49, -0.37]$. The largest effect occurs at one lead and gradually decays in both directions. This pattern explains why the calculated correlation coefficients involving output in Table 1, which involve first differences of the data, are positive at leads 2 and 3.

ii. *Security Prices*. Using equation (11) and the fact that τ follows a two-state Markov process, the price of an equity claim can be simplified to

$$(20) \quad q(\tau_i) = \frac{(1 - \alpha)\beta(1 - \gamma(\tau_i))}{(1 - \beta)(1 - \beta\delta)} \\ \times y[(1 - \beta\pi_i)g(1 - \tau_i) \\ + \beta(1 - \pi_j)g(1 - \tau_j)]$$

for $i, j = 0, 1$ and $i \neq j$. For the case where $\pi_0 = \pi_1 = \pi$, equation (20) along with the definition of $\gamma(\tau)$ implies that the sign of $q(\tau_0) - q(\tau_1)$ is the same as the sign of $[g(1 - \tau_0) - g(1 - \tau_1)][1 - \beta - \alpha\beta^2(1 - \pi)(g(1 - \tau_0) + g(1 - \tau_1))]$. When $\pi > 1/2$,

the first bracketed expression is positive, while the second depends on the values of the relevant parameters. For π approaching unity, the second term is positive. The reason that the sign of $q(\tau_0) - q(\tau_1)$ is ambiguous is because equity prices are affected through two channels. With lower taxes and persistence of the tax process, individuals expect higher dividends, driving equity prices up. However, a low tax state implies lower current consumption, which tends to lower equity prices.

Of more interest are the correlation coefficients between equity prices and tax rates. Unlike those for output, their sign is not so clear-cut. However, it generally does not take much persistence to yield a negative correlation between q_{t+j} and τ_t for $j = 0, 1, 2, 3$. This result occurs because, when tax rates persist, consumption will rise over time when taxes are low and so too will equity prices. Therefore, low tax rates are more likely to be associated with high security prices. The vector of correlation coefficients for the parameter values used in the previous section is $[-0.74, -0.84, -0.40, -0.32, -0.24, -0.16]$, with the peak occurring at lead two. For less persistence, $\pi = 0.8$, a similar pattern was obtained, while for $\pi = 0.55$, all but the contemporaneous and first lagged correlation coefficient were negative. Further, since effective tax rates are inversely related to inflation, the model is capable of generating the negative correlation between stock prices and inflation that is commonly observed in the United States.

The results of this section can be viewed as a stochastic analogue of the results in Brock and Turnovsky (1981) in the case where firms use equity financing. In their article, share value would rise during an adjustment from a high to low tax rate, while equity prices would initially fall and then rise as consumption increased to its new steady-state value. However, in their model the steady-state value of equity prices only depends on corporate taxes through their effect on the marginal utility of consumption. This feature is due to the simple nature of the process for dividends in which dividends do not vary with corporate tax rates.

iii. *The After-Tax Real Rate Of Interest.* Calculating the equilibrium after-tax risk-free real rate of interest, ρ , is accomplished by considering the price, p^R , of a tax-free bond, B , that yields one unit of consumption next period. This is done by adding $p_t^R B_{t+1}$ to the left-hand side of (3); and B_t to the definition of wealth. The resulting first-order condition with respect to B_{t+1} is

$$(21) \quad E_t \lambda_{t+1} = \lambda_t p_t^R.$$

Equation (21) implies that an individual is indifferent at the margin between sacrificing p_t^R units of wealth today for one unit of wealth next period.

Using the expressions for consumption and investment and the fact that $1/[1 - \gamma(\tau_t)] = 1 + \alpha\beta g(1 - \tau_t)$ implies that

$$(22) \quad 1 + \rho_t = \frac{1}{p^R} = \frac{1/c_t}{\beta E_t(1/c_{t+1})} \\ = \frac{1 + \alpha\beta g(1 - \tau_t)}{\beta E(1 + \alpha\beta g(1 - \tau_{t+1}))} \\ \times \gamma(\tau_t)^\alpha y_t^{\alpha-1}.$$

For the case where τ follows the process given by (13a) and (13b) and where taxes show persistence ($\pi_0, \pi_1 > 1/2$), it can be shown that the after-tax real rate of interest is higher when a low tax state is realized.⁹ This rise in real rates occurs because a lowering of the tax rate indicates that capital will be more valuable in the future and that there will be more future output. Individuals will, therefore, value a unit of future

⁹The proof of this relies on the facts that if $\pi_0, \pi_1 > 1/2$ then $g(1 - \tau_0) > g(1 - \tau_1)$ and that

$$\frac{1 + \alpha\beta g(1 - \tau_0)}{\beta(1 + \alpha\beta E g(1 - \tau_{t+1} | \tau_t = \tau_0))} \\ > \frac{1 + \alpha\beta g(1 - \tau_1)}{\beta(1 + \alpha\beta E g(1 - \tau_{t+1} | \tau_t = \tau_1))}.$$

The proof of the latter inequality involves some cumbersome algebra and is omitted.

wealth by less, causing p^R to fall and $1 + \rho_t$ to rise. Alternatively, individuals will wish to accumulate more capital. In order to induce lower consumption, the real rate of interest must rise.

Regarding correlation coefficients, the contemporaneous and lagged correlations are all negative, while the lead correlations are generally positive (the exception being $\text{Cov}(\hat{\pi}_t, \ln(1 + \rho_{t+1}))$ when $\pi \geq 0.9$).

iv. *A Comparison with the Data.* Because the time-series of output, real interest rates, and stock prices appear to be nonstationary, correlations between the log of the effective tax rate on capital and the difference of the logs of the various series were used. In Table 1, the actual correlation coefficients are compared with those generated by the model. The results are not overly encouraging, but it would be presumptuous to believe that the extraordinarily simple model in this section would closely mimic the data. In addition to the model's simplicity, the actual data, particularly the effective tax-rate series and the *ex ante* real rate, suffer from measurement error. A more thorough analysis, which is beyond the scope of this paper, would incorporate taxes into a more detailed model of the economy to see if including taxes would significantly improve the model's ability to replicate the data.

For the model in this section the best results occur for output, where the contemporaneous and first lead correlation coefficient are not too different from those predicted by the model. The model's results for *ex ante* after-tax real rates exhibit some of the same sign patterns as the data, while the actual data for equity prices do not match the model at all.

III. Investment Tax Credits

As claimed earlier in the paper, the simplified treatment of the tax structure does not change the basic message, that uncertainty should be explicitly modeled. Also, as shown in Section IV, explicitly modeling uncertainty, as opposed to assuming perfect foresight, tends to reduce the range of values attained by the capital stock for both the case of a tax on production and an

investment tax credit. To make things simple, the effects of a stochastic investment tax credit are examined separately rather than studying both production taxes and the investment tax credit simultaneously.

The change in the model is straightforward. Lump-sum government transfers are now negative and equal $-\theta_t K_{t+1}$, where θ_t is the investment tax credit at time t . Its value is known at time t , but its future values are uncertain. Each individual receives $\theta_t k_{t+1}$, and this term is added to the right-hand side of the individual's budget constraint. That is, instead of (3), we now have

$$(3') \quad c_t + k_{t+1} + q_t s_{t+1} \leq w_t + \theta_t k_{t+1}$$

and equation (6b) would now be

$$(6b') \quad \beta E D_1 V(k', K', s', \theta') = \lambda(1 - \theta).$$

With exponential production and log utility, the proportion of output invested is given by

$$(9') \quad \eta(\theta) = \frac{\alpha \beta g(\theta)}{1 + \alpha \beta g(\theta)}$$

where $g(\theta)$ is determined by the recursive relationship

$$(10') \quad g(\theta) = \frac{1}{1 - \theta} [1 + E \alpha \beta g(\theta')].$$

From (10') one observes that the current investment tax credit, unlike the current profits tax, directly affects investment, since it influences the relative value of k_{t+1} . Therefore, even if θ were distributed independently and identically, investment would be influenced by realizations of θ .

It is straightforward but tedious to show that, for reasonable parameter values when θ follows a two-state Markov process and when tax credits show persistence, investment is higher when θ is high. Further, since investment decisions also involve expectations of future productivity, which depend on future levels of capital, the persistence of the process generating θ will be important. Also, the more persistent θ is,

TABLE 2

Time Period	Tax Rate	$\gamma(\tau_t)$ Under Perfect Foresight	k_{t+1} Under Perfect Foresight	$\gamma(\tau_t)$ Under Uncertainty	k_{t+1} Under Uncertainty
0			0.095		0.095
1	0.50	0.167	0.076	0.18	0.082
2	0.50	0.167	0.071	0.18	0.078
3	0.50	0.167	0.069	0.18	0.077
4	0.50	0.167	0.068	0.18	0.077
5	0.50	0.167	0.068	0.18	0.076
6	0.50	0.167	0.068	0.18	0.076
7	0.50	0.169	0.069	0.18	0.076
8	0.50	0.180	0.074	0.18	0.076
9	0.50	0.250	0.105	0.18	0.076
10	0.25	0.250	0.118	0.24	0.102
11	0.25	0.250	0.123	0.24	0.112
12	0.25	0.250	0.124	0.24	0.116
13	0.25	0.250	0.125	0.24	0.117
14	0.25	0.250	0.125	0.24	0.117
15	0.25	0.245	0.123	0.24	0.118
16	0.25	0.231	0.115	0.24	0.118
17	0.50	0.167	0.081	0.18	0.089
18	0.50	0.167	0.072	0.18	0.080
19	0.50	0.167	0.069	0.18	0.078
20	0.50	0.167	0.068	0.18	0.077

the greater the difference in $\eta(\theta)$ when tax credits are high as opposed to low. Again, this result is similar to one reported in Bizer and Judd (1988). Therefore, taking account of the exact nature of the generating mechanism for θ is necessary for analyzing economic behavior.

IV. A Numerical Example

In this section, a direct comparison is made between capital stock accumulation under perfect foresight and situations where agents behave under uncertainty. To perform the experiment, 20 realizations of tax rates and investment tax credit rates were generated using a random number generator and the actual transition probabilities associated with the effective tax rate series given in Hulten and Robertson (1982).¹⁰ The

parameter values used were $\alpha = 1/3$, $\beta = 1$, $\tau_0 = 1/4$, $\tau_1 = 1/2$, and $\pi_0 = \pi_1 = 0.9$. For investment tax credits, $\theta_0 = 0.10$ and $\theta_1 = 0$. The realizations and resulting levels of the capital stock are given in Tables 2 and 3. The starting value of the capital stock for the first experiment is the value that would result if $\tau = 0.375$ (the average value of the tax rate) for all time and, for the second experiment, the value that would result if $\theta = 0.05$ for all time. In calculating the values under the assumption of perfect foresight, it was assumed that $g(1 - \tau_{21}) = 0.60$ and $g(\theta_{21}) = 1.765$. These assumptions are important regarding the solution for the last period's capital stock, but the terminal values have almost no effect on the numbers reported in Tables 2 and 3.

From Table 2, it is clear that the capital stock under uncertainty behaves in a quantitatively different manner than it does under perfect foresight. The movements in capital

¹⁰For the production tax, the exact procedure was to look at the Hulten and Robertson effective tax rate series as a realization of a two-state first-order Markov process and use the calculated sample transition probabilities and sample means. Then, 20 random numbers between zero and one were generated. It was assumed that the initial tax rate was high and that a number

between 0 and 0.1 implied a change in the tax rate, while a number between 0.1 and 1.0 implied that the tax rate remained at its previous value. For the investment tax credit, no actual series was available. It was assumed that $\pi_0 = \pi_1 = 0.9$ for this process as well.

TABLE 3

Time Period	Investment Tax Credit	$\eta(\theta_t)$ Under Perfect Foresight	K_{t+1} Under Perfect Foresight	$\eta(\theta_t)$ Under Uncertainty	K_{t+1} Under Uncertainty
0			0.208		0.208
1	0.10	0.370	0.219	0.368	0.218
2	0.10	0.370	0.223	0.368	0.221
3	0.10	0.370	0.225	0.368	0.223
4	0.10	0.370	0.225	0.368	0.223
5	0.10	0.370	0.225	0.368	0.223
6	0.10	0.370	0.225	0.368	0.223
7	0.10	0.369	0.224	0.368	0.223
8	0.10	0.366	0.222	0.368	0.223
9	0.10	0.357	0.216	0.368	0.223
10	0	0.333	0.200	0.335	0.203
11	0	0.333	0.195	0.335	0.197
12	0	0.333	0.193	0.335	0.195
13	0	0.333	0.193	0.335	0.194
14	0	0.335	0.193	0.335	0.194
15	0	0.338	0.195	0.335	0.194
16	0	0.346	0.201	0.335	0.194
17	0.10	0.370	0.217	0.368	0.213
18	0.10	0.370	0.223	0.368	0.220
19	0.10	0.370	0.224	0.368	0.222
20	0.10	0.370	0.225	0.368	0.223

tend to be smoothed out by uncertainty, since there is always some positive probability that next period's tax rate will be different from today's. Also, under perfect foresight, agents respond one period sooner to tax rate changes than do agents who are unsure of the value of next period's tax rate.

This difference in behavior would also occur if taxes were independently and identically distributed. Under uncertainty, the level of the capital stock would remain at 0.095, independent of the actual realizations of taxes, while with perfect foresight the capital stock would respond considerably. Therefore, if one is to predict accurately how agents will respond to tax rate changes, one must carefully consider the forecasting problem facing agents. Doing so requires an explicit stochastic treatment of the problem.

Regarding Table 3 one observes that the relative range of capital values is slightly smaller under uncertainty and that agents do not alter their behavior prior to new realizations of the investment tax credit. However, unlike the case with a tax on production, one-period movements in the capital stock can be somewhat sharper under uncertainty.

V. Summary

This article analyzes the effects of taxes on production in a simple stochastic growth model. The paper represents an advance since it treats tax rates as inherently stochastic. As shown, the actual process generating taxes is an important determinant in understanding how the economy will behave with respect to particular realizations of tax rates.

One might also wish to explore the interaction between tax changes and nominal magnitudes such as inflation and the nominal interest rate. Extending the model to incorporate money via a cash-in-advance constraint is not a problem. The basic solution governing the real side of the economy is unchanged if the cash-in-advance constraint is only on consumption and if one rules out precautionary demands for cash. This is easily done so long as monetary growth is not overly deflationary. The solution is only slightly changed if the cash-in-advance constraint includes capital as well. In these cases, a decrease in money growth causes capital accumulation and an increase in output. For given money growth, the price level falls, and the economy experiences

lower inflation. Since little is to be gained through these additions, the paper concentrates on real economic variables.

APPENDIX

This Appendix provides a more formal description of equilibrium and shows that the solution in the text is an equilibrium: $\{p_t^*, q_t^*, c_t^*, x_t^*, s_t^*, k_t^*, \pi_t^*, T_t^*\}_{t=0}^\infty$ is an equilibrium given the tax process $\{\tau_t^*\}_{t=0}^\infty$ if

(A1) $\{c_t^*, s_t^*, x_t^*\}$ solves the consumer's problem

$$\max E \left[\sum_{t=0}^{\infty} \beta^t u(c_t) \right]$$

subject to

$$\begin{aligned} c_t + x_{t+1} + q_t^* s_{t+1} &\leq (q_t^* + \pi_t^*) s_t + p_t^* x_t + T_t^* \\ x_0 &= k \\ s_0 &= 1 \end{aligned}$$

where $u(c_t) = \ln c_t$;

(A2) $\{k_t^*\}$ solves the firms problem

$$\max (1 - \tau_t^*) f(k_t) - p_t^* k_t$$

where $f(k_t) = k_t^\alpha$ and $k_t \geq 0$;

(A3) $\pi_t^* = (1 - \tau_t^*) f(k_t^*) - p_t^* k_t^*$;

(A4) $k_t^* = x_t^*$;

(A5) $s_t^* = 1$;

and

(A6) $T_t = \tau_t f(k_t^*)$

where τ follows a first-order Markov process is satisfied. Substituting the postulated solutions $c = [1 - \gamma(\tau)]k^\alpha$ and $k' = \gamma(\tau)k^\alpha$ into (7b) yields

$$\alpha \beta E \frac{1}{\gamma(\tau)} \frac{1 - \tau'}{1 - \gamma(\tau')} = \frac{1}{1 - \gamma(\tau)}.$$

The solution to this nonlinear first-order difference equation is satisfied for

$$\gamma(\tau) = \frac{\alpha \beta g(1 - \tau)}{1 + \alpha \beta g(1 - \tau)}$$

where $g(1 - \tau_t) = E_t[(1 - \tau_{t+1})[1 + \alpha \beta E_{t+1}[(1 - \tau_{t+2}) \times [1 + \alpha \beta E_{t+2}[\dots]]]]]$

or

$$g(1 - \tau) = E\{(1 - \tau')[1 + \alpha \beta g(1 - \tau')]\}.$$

This solution satisfies all six equilibrium conditions.

REFERENCES

- Abel, Andrew B. and Blanchard, Oliver, "An Intertemporal Model of Saving and Investment," *Econometrica*, May 1983, 51, 675-92.
- Barro, Robert J., "On the Determination of Public Debt," *Journal of Political Economy*, October 1979, 87, 940-71.
- Becker, Robert A., "Capital Income Taxation and Perfect Foresight," *Journal of Public Economics*, 1985, 26, 147-67.
- Bizer, David and Judd, Kenneth L., "Capital Accumulation, Risk, and Uncertain Taxation," unpublished manuscript, March 1988.
- Boyd, John H. III, "Symmetries, Equilibria and the Value Function," Rochester Center for Economic Research, Working Paper No. 62, December 1986.
- Brock, William A., "Asset Prices in a Production Economy," in John J. McCall, ed., *The Economics of Information and Uncertainty*, Chicago: University of Chicago Press, 1979, 1-43.
- and Turnovsky, Stephen J., "The Analysis of Macroeconomic Policies in Perfect Foresight Equilibrium," *International Economic Review*, February 1981, 22, 179-209.
- Coleman, Wilbur John II, "Equilibrium in an Economy with Capital and Taxes on Production," unpublished manuscript, May 1989.
- Danthine, Jean Pierre and Donaldson, John B., "A Note on the Effects of Capital Income Taxation on the Dynamics of a Competitive Economy," *Journal of Public Economics*, November 1985, 28, 255-65.
- Goulder, Lawrence H. and Summers, Lawrence H., "Tax Policy, Asset Prices, and Growth: A General Equilibrium Analysis," National Bureau of Economic Research, Working Paper No. 2128, January 1987.
- Hall, Robert E., "The Dynamic Effects of Fiscal Policy in an Economy with Foresight," *Review of Economic Studies*, Jan-

uary 1971, 38, 229-44.

Hulten, Charles R. and Robertson, James W., "Corporate Tax Policy and Economic Growth: An Analysis of the 1981 and 1982 Tax Acts," Discussion Paper, Urban Institute, December 1982.

Lucas, Robert E., Jr., "Econometric Policy Evaluation: A Critique," in K. Brunner and A. Meltzer, eds., *The Phillips Curve*

and the Labor Market, Carnegie-Rochester Conference Series in Public Policy, Vol. 1, Amsterdam: North Holland, 1976.

_____ and **Stokey, Nancy L.**, "Optimal Fiscal and Monetary Policy in an Economy Without Capital," *Journal of Monetary Economics*, July 1983, 12, 55-94.

Sargent, Thomas J., *Macroeconomic Theory*, New York: Academic Press, 1979.

Deposit Insurance, Risk, and Market Power in Banking

By MICHAEL C. KEELEY*

A fixed-rate deposit insurance system provides a moral hazard for excessive risk taking and is not viable absent regulation. Although the deposit insurance system appears to have worked remarkably well over most of its 50-year history, major problems began to appear in the early 1980's. This paper tests the hypothesis that increases in competition caused bank charter values to decline, which in turn caused banks to increase default risk through increases in asset risk and reductions in capital. (JEL 600)

It has long been recognized that a fixed-rate deposit insurance system, such as the Federal Deposit Insurance Corporation's (FDIC's), or the Federal Savings and Loan Insurance Corporation's (FSLIC's) can pose a moral hazard for excessive risk taking. The reason is that banks or thrifts can borrow at or below the risk-free rate by issuing insured deposits and then investing the proceeds in risky assets with higher expected yields.

As Robert C. Merton (1977) has shown, deposit insurance can be viewed as a put option on the value of a bank's assets at a strike price equal to the promised maturity value of its debt. Under a fixed-rate system, banks potentially can transfer wealth from the insuring agency, and, absent regulation, banks seeking to maximize the value of their equity will maximize the value of the put by

increasing asset risk and/or minimizing invested capital relative to assets.

Empirical research, however, does not seem to show that banks in general maximize the put option value. For one thing, many banks hold substantially more capital than the required amounts (Michael C. Keeley, 1988) and for another, researchers have found that for many banks, the value of the deposit insurance option is less than its price (Allan J. Marcus and Israel Shaked, 1984; Ehud Ronn and Avinash K. Verma, 1986; George Pennacchi, 1987), assuming that at the expiration of the option insolvent banks are closed. Moreover, for most of its 50-year history, the insurance system has been characterized by low failure rates and low payouts—just the opposite of what might be expected if banks were maximizing the value of the put option successfully.

Recently, bank and thrift failures and deposit insurance payouts have reached record highs (see Chart 1); the FSLIC has liabilities far in excess of its assets, and even the FDIC faces threats to its solvency. Although many have argued that these recent problems are in part due to the moral hazard of deposit insurance, the question is why it has taken 50 years for major problems to arise.

One explanation (Arnold Kling, 1986) is that the recent episode simply reflects an increasingly risky economy, which in turn has increased the risk of bank portfolios. In the last few years, whole sectors and regions of the national and even the world economy have encountered serious downturns that

*Vice President, Cornerstone Research, 1000 El Camino Real, Menlo Park, CA 94025. Much of the research in this paper was conducted while the author was a research officer at the Federal Reserve Bank of San Francisco. However, opinions expressed herein are those of the author and do not necessarily reflect the view of the Federal Reserve Bank of San Francisco, the Board of Governors of the Federal Reserve System, or Cornerstone Research. An earlier version of this paper was presented at Garn Institute of Finance's academic symposium on deposit insurance. Comments from William Beaver, Jack Beebe, Barbara Bennett, Mark Flannery, Christopher James, Ed Kane, Stuart Myers, Randall Pozdena, Anthony Saunders, and two anonymous referees are greatly appreciated. Alice Jacobson provided expert research assistance. The usual caveats apply, however.

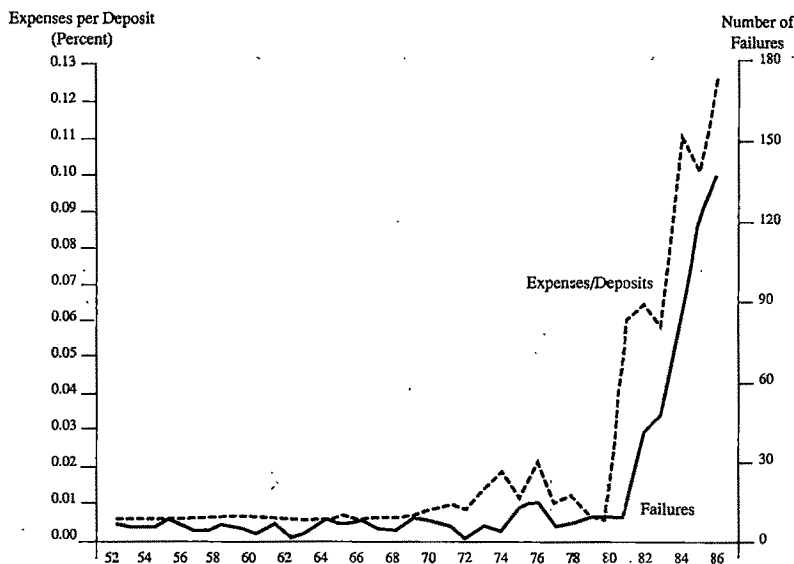


CHART 1. DEPOSIT INSURANCE EXPENSES PER DOLLAR OF DEPOSITS AND BANK FAILURES

have affected the values of bank and thrift assets. Similarly, interest rates have become more volatile, increasing the riskiness of banks', and especially thrifts', portfolios.

The rise in bank and thrift failures in recent years also may reflect the secular decline in capital-to-asset ratios over the past two decades. As Chart 2 shows, both market and book capital ratios of the 25 largest bank holding companies have fallen well below their levels in the mid-1950's, when only a handful of banks and thrifts failed each year, as opposed to several hundred per year recently. Moreover, beginning in about 1974, market values of the 25 largest bank holding companies in the aggregate fell below book capital ratios.

There are two reasons why declining capital ratios could lead to an increased rate of bank failures. First, lower capital, holding asset risk constant, leads to less protection against failure. Second, as shown in Frederick T. Furlong and Keeley (1987, 1989), lower capital ratios increase the incentive for banks to increase asset risk. Thus, even if overall risk in the economy did not increase, banks would have a greater incentive to increase asset portfolio risk due to the decline in capital ratios.

There is little doubt that increased risk in the economy and declining capital ratios have had a lot to do with the increase in bank and especially thrift failures in recent years. But these developments do not explain why banks and thrifts allowed bankruptcy risk to increase. After all, depository institutions have considerable control over the riskiness of their asset portfolios and perhaps even more control over their capital ratios. Thus, these explanations beg the question of why capital ratios behaved as they did.

Specifically, why did banks on average hold so much capital during the 1950's and early 1960's, and why did capital ratios fall during the 1960's and 1970's? It seems difficult to pinpoint any explicit regulatory changes that would have made it easier for banks to increase default risk, and banks had access to fixed-rate deposit insurance throughout the period.¹ A similar puzzle

¹Although the percentage of deposits explicitly covered by deposit insurance has increased, in theory, even partial deposit insurance coverage provides an incentive for banks to minimize capital and maximize asset risk (as long as they can share losses with the deposit insurance fund).

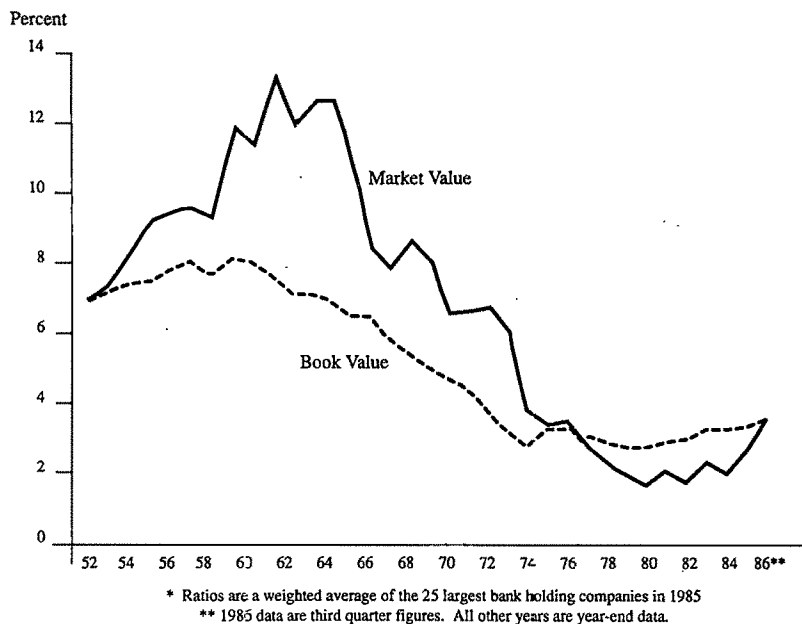


CHART 2. CAPITAL-TO-ASSET RATIOS, MARKET AND BOOK VALUES

arises in trying to explain the cross-sectional variation in bank capital ratios. Some banks, for example, hold much more capital than others and much more than regulators require.

This paper argues that one explanation for these apparent puzzles involves differences and changes in the degree of competition faced by banks. In the 1950's and even early 1960's banks partially were protected from competition by a variety of regulatory barriers. For example, chartering was very restrictive (Sam Peltzman, 1965) until the mid-1960's when James Saxton, then comptroller of the currency, greatly liberalized it (Keeley, 1985a, b). Moreover, some banks were protected by various state laws that limited or prohibited branching, multi-bank holding company, and interstate bank expansion. However, these laws have been greatly liberalized over the last few years, possibly eroding banks' charter values. Likewise, deposit rate deregulation may have diminished charter values by increasing competition, especially for institutions in protected local markets that had been relying on nonprice service competition to attract funds. In addition, beginning in the

early 1980's, thrifts were given expanded powers that enabled them to compete more fully with banks. Finally, many argue that changes in technology have increased the competition that banks face from nonbank financial firms, such as investment banks, brokerage firms, and insurance companies. Such developments as money market mutual funds, cash management accounts, and increased use of commercial paper have all made competitive inroads in banks' traditional product markets.

Increased competition may have reduced banks' incentives to act prudently with regard to risk taking. In fact, the evidence in Chart 2 of declining market values (which would reflect capitalized charter values) relative to book values (which would not) suggests that bank charter values were declining. In the 1950's and early 1960's, regulatory restrictions on entry and competition made bank charters valuable. With valuable charters as assets, banks had an incentive not to risk failure since the owners of the banks cannot sell the charter once the bank is declared insolvent. Instead, a bank that was insolvent on a book-value basis still had a valuable charter that the FDIC could sell

in a purchase and assumption (P&A). (In calculating primary book capital, intangible assets, generally representing the excess of the purchase price of assets that had been acquired over their book value, are subtracted from capital.) This may explain why P&A's typically are less costly than liquidations. Also, banks would apparently be willing to overpay for deposit insurance if it were needed to obtain and maintain a valuable charter.

The possession of a valuable charter thus made it difficult for banks to shift losses to the FDIC, and its potential loss in essence created a regulatory bankruptcy cost from the point of view of bank owners. This is especially so since regulators focus on book value when assessing a bank's solvency, not market value. Thus, the gains from feasible increases in risk taking would be offset by the diminished expected value of the charter. As a result, a bank will not have an incentive on the margin to increase default risk (either through reducing capital relative to assets or increasing asset risk) as long as the expected loss of the charter exceeds the gain to the bank of the enhanced value of the deposit insurance put option. Moreover, regulation limited the feasible increases in risk taking so as to prevent banks from potentially imposing losses that would have exceeded their charter values. This idea is established formally below using a state preference model.

In the empirical analysis below a simultaneous equations model of bank risk taking and charter value is developed and estimated. Changes in the laws governing branching, multibank holding company expansion, and interstate entry are used to identify the model statistically. Over the last 20 years or so, these anticompetitive laws have been liberalized greatly, and there are virtually no cases of states increasing their stringency (Dean F. Amel and Daniel G. Keane, 1987). Although the liberalization of these laws is not necessarily the most important factor in increasing the degree of bank competition, it is an easily observed exogenous factor with respect to bank risk taking. Thus, changes in these laws over time provide an opportunity to examine their

influence on market power in banks and whether exogenous variations in market power are related to bank risk taking.

This paper first examines the relationship between changes in regulatory entry barriers and the market power of banks in order to create an instrument for charter value. Then the relationship between market power and risk taking is estimated. This paper employs James Tobin's q , as suggested by Eric Lindenberg and Stephen Ross (1981), as a measure of a bank's market power (monopoly rents). Two measures of bank risk are then related to exogenous variations in q : the market-value capital-to-asset ratio and the interest cost on large, uninsured CD's. I find that q appears to be a useful proxy for market power and that banks with greater market power hold more capital and pay lower rates on CD's.

The remainder of the paper is organized as follows. In Section I, a state preference model is employed to show how market power can affect bank risk taking and how it can be measured. Section II presents the empirical results. Finally, Section III contains a summary and conclusions.

I. Theoretical Framework

Below, a state preference model is used to develop the major results.² The model described below closely follows that presented in Furlong and Keeley (1989). Marcus (1984) has developed similar results using an options model, but the state preference model clarifies the conditions under which a bank can benefit from increasing default risk and also illustrates the relationship between charter value and Tobin's q .

A two-period (current and future period), two-state model is used where P_1 and P_2 are the current values of a dollar payment in future states 1 and 2, respectively. (Thus, the risk-free interest rate is $1/(P_1 + P_2) - 1$).

²State preference models have been widely used in the analysis of banking and deposit insurance—see John H. Kareken and Neil Wallace (1978), William F. Sharpe (1978), Uri Dothan and Joseph Williams (1980), and Furlong and Keeley (1987, 1989).

The state prices are assumed to be exogenously given. To fund its assets, a bank uses an initial capital of C_0 dollars and issues deposits with a current value of D_0 dollars. Initially, the bank is assumed to issue risk-free deposits that pay off \$1 in each state, although this assumption is relaxed later to allow for the issuance of risky deposits. Also, it is assumed that initially the bank is not insured, although this assumption is relaxed later too. The bank invests in an asset security A that pays A_1 dollars in state 1 and A_2 dollars in state 2.

The current value of the bank's equity, V_0 , is found by valuing the various cash flows at the state prices. It is assumed that the bank can acquire security A at a price P_A and issue deposits at a price P_d . Thus, the bank can purchase $(C_0 + D_0)/P_A$ units of A. The current value of the bank's equity, V_0 , is the value of the cash flows from the assets acquired minus those of the liabilities issued:

$$(1) \quad V_0 = [(C_0 + D_0)/P_A][P_1A_1 + P_2A_2] - (D_0/P_d)(P_1 + P_2).$$

If the bank is competitive in both the asset market and the deposit market, then $P_A = P_1A_1 + P_2A_2$ and $P_d = P_1 + P_2$, and equation (1) above reduces to

$$(2) \quad V_0 = C_0.$$

If the bank is not insured, the value of its equity is independent of its (*ex ante*) risk taking. The reason is that if the bank were to default in state 1 and not meet its promised obligations to depositors, the depositors would demand sufficiently higher payments in state 2 so as to leave the costs unchanged at P_d .

However, with deposit insurance, depositors would not demand a higher payment in state 2, because the insuring agency would pay them the difference between the promised obligations and the asset value in the event of bankruptcy. But with fixed-rate underpriced deposit insurance with a premium, assumed for expositional purposes to be zero, banks pay less than the promised

amount if bankruptcy occurs in state 1, and only the promised amount in state 2. If the bank were to go bankrupt in state 1, then the future value of the bank's equity in state 1 is zero, and its current equity value is

$$(3) \quad V_0 = (C_0 + D_0)(P_2A_2)/P_A - P_2D_0/P_d > C_0.$$

The current value of the bank's equity when bankruptcy would occur in state 1 (that is, when bankruptcy is possible) equals the value of the excess of its deposit obligations over asset returns (the option value of deposit insurance) I_0 , plus its invested capital, C_0 . That is,

$$(4) \quad V_0 = I_0 + C_0$$

where

$$(5) \quad I_0 = D_0P_1/P_d - (C_0 + D_0)P_1A_1/P_A$$

and $I_0 > 0$.

As is well known, increasing capital, holding constant deposits, reduces the value of the deposit insurance option, and hence equity, and increasing asset risk (increasing the payment in state 2 while reducing that in state 1 so as to hold the price of the asset constant) increases the option value.³ Thus, the problem facing bank regulators is to constrain banks' incentives to reduce capital relative to assets and to increase asset risk. The puzzle is why regulators apparently succeeded throughout much of the last 50 years but in recent years apparently have failed.

³Since $I_0 > 0$, the value of the bank exceeds its initial capital investment. Thus,

$$dI_0/dC_0|D_0 = -P_1A_1/P_A < 0$$

and

$$dI_0/dA_2|P_A = -(C_0 + D_0)(P_1/P_A)(dA_1/dA_2) > 0$$

since

$$dA_1/dA_2 < 0.$$

A. Charter Values

If banks can operate only with charters that are limited in supply, banks may be able to acquire assets at below-market prices (that is, bank loans would earn higher risk-adjusted rates than would market securities) and/or they may be able to make below-market-value payments on deposits (that is, deposits would pay below the risk-adjusted rate). Bank charters have been made valuable by limiting their supply and by protecting banks through various regulations that limited interbank competition as well as competition by nonbank firms.

In the model below, banks are assumed to be insured (at zero cost)⁴ but face periodic examinations. At the end of the period, if the bank is insolvent (that is, its assets, not including the charter value, are less than deposit obligations), equity holders receive nothing, depositors receive their promised obligations, and the insurance agency receives the bank's assets, including the charter. If the bank is solvent, however, the bank retains its charter value and continues to operate for another period.

If a bank chooses capital and asset risk so that it will not default in either state, the current value of the bank's equity is

$$(6) \quad V_0 = (C_0 + D_0)(P_1A_1 + P_2A_2)/P_A \\ - D_0(P_1 + P_2)/P_d + X_0$$

where

$$(7) \quad X_0 = P_1X_1 + P_2X_2$$

in which P_1X_1 is the current value of the charter to operate one more period if state 1 occurs, and P_2X_2 is the current value of the charter to operate one more period if state 2 occurs. Thus, the bank must balance the gains from increased risk taking (I_0) with the loss of the charter value if bankruptcy occurs (P_1X_1). The bank will risk

bankruptcy only if

$$(8) \quad I_0 > P_1X_1.$$

Consider a bank just on the verge of insolvency in state 1 (that is, when a marginal increase in asset risk or reduction in capital would cause bankruptcy). The value of such a bank is

$$(9) \quad V_0 = (C_0 + D_0)(P_2A_2)/P_A \\ - D_0P_2/P_d + X_0.$$

However,

$$(10) \quad dV_0/dD_0|C_0 = -P_1X_1 < 0.$$

That is, a marginal increase in deposits holding capital constant, which causes bankruptcy in state 1, in turn causes the bank to lose the value of its charter if state 1 occurs, P_1X_1 . Similarly,

$$(11) \quad dV_0/dA_2|P_A, C_0 = -P_1X_1 < 0.$$

Thus, a bank initially at a position where solvency is guaranteed in both states will not have an incentive on the margin to increase risk either through increases in leverage (that is, increases in deposits holding capital constant) or increases in asset risk.⁵ This remains true throughout the region where $P_1X_1 > I_0$.

Although valuable bank charters do not obviate the need for bank regulation because I_0 is unbounded in the absence of

⁵Marcus (1984) shows that dV_0/dD_0 can be positive for banks with low charter values and negative for banks with high charter values using an options pricing formula. Although he argues that $dV_0/d\sigma$ (the equivalent of dV_0/dA_2) can be negative, his figure 2 implies that, for banks with high charter values, $dV_0/d\sigma$ is positive. Moreover, in his model, it is unclear what determines the critical value of whether a marginal increase in default risk will be beneficial.

In contrast, the state preference model shows that the choice is one of balancing the gains in the option value of deposit insurance with loss of the expected value of the charter. Marginal increases in default risk will not benefit the bank until default risk is sufficiently high so that $I_0 = P_1X_1$.

⁴Similar results are obtained if deposit insurance has a positive cost not related to default risk. The assumption of zero cost is employed to simplify the analysis.

regulation, they make the regulator's job much easier. If a bank's capital and asset risk were initially set so that solvency were assured in both states, a large discrete increase in asset risk or reduction in capital sufficient to make $I_0 > P_1 X_1$ would be required if the value of the bank's equity were to increase. Since such large discrete changes presumably would be easy to detect, banks would be discouraged from trying to increase default risk in the first place, and regulators would not need to be concerned with small changes in asset risk or capital.

B. Market Power

An uninsured bank that has market (monopoly) power in its asset market can make positive net present value loans. That is, the loans' future payoffs, when valued at the exogenously given state prices, exceed their current cost. For a such a bank,

$$(12) \quad (P_1 A_1 + P_2 A_2) / P_A = \varepsilon > 1.$$

However, the bank does not face an inexhaustible supply of such loan opportunities, and as assets A_0 (which equal $C_0 + D_0$) increase, ε , the ratio of cash flows from an asset to its price diminishes. (That is, $\varepsilon = \varepsilon(A_0)$ and $\varepsilon' < 0$.) Assuming such a bank maximizes its net-of-capital investment value, it will expand until the current value of its marginal revenue equals the marginal cost of its deposits (which is 1). That is, $\varepsilon(1+n)=1$, where n is the elasticity of ε with respect to A_0 . (For an uninsured bank, this condition holds regardless of whether bankruptcy occurs, assuming no bankruptcy costs, since the costs of deposits are unaffected by the risk of bankruptcy.) Similar conditions hold for market power on the deposit side.⁶

For banks that have market power in either the asset or deposit market, as suggested originally by George Stigler (1964) and later and more formally by Lindenberg

and Ross (1981), Tobin's q is an ideal measure of market power. In this paper, Tobin's q is defined as the current market value of a firm's assets (the market value of its equity plus debt) divided by their current cost to a firm. The reason that q is an ideal measure of monopoly rents is that the capitalized value of such rents, whether they arise from market power in the asset market, deposit market, or both, will be reflected in the market value of the firm's equity, and thus assets, but not in the costs of acquired assets. The reasons for q 's superiority as a measure of market power are spelled out in more detail in Michael Salinger (1984) and Michael Smirlock et al. (1984). To see why Tobin's q is a measure of market power in the above model, note that q is given by

$$(13) \quad q = \frac{[(C_0 + D_0)\varepsilon - D_0 + X_0] + D_0}{C_0 + D_0} \\ = \varepsilon + X_0 / (C_0 + D_0).$$

(The terms in brackets [] represent the market value of the bank's equity to which the current value of debt D_0 is added.) Equation (13) shows that q is equal to the current plus future degree of market power as reflected in current and future ε . For a competitive firm, $\varepsilon = 1$ and $X_0 = 0$, but for a firm with market power, as discussed above, $\varepsilon > 1$ and $X_0 > 0$. Thus, an uninsured bank with no market power in either the asset or deposit market would have a q of 1.⁷

II. Empirical Evidence

The data used to estimate the model are from several sources. The bank holding company data are from the Compustat bank

⁷For a bank with market power on the deposit side as well,

$$q = \varepsilon + \frac{D_0}{C_0 + D_0}(1-f) + \frac{X_0}{C_0 + D_0} \quad \text{where } f < 1$$

is inversely related to the degree of market power on the deposit side (see footnote 6). Thus, a bank with market power on the deposit side also will have a q greater than one.

⁶For a bank with market power on the deposit side, $(P_1 + P_2) / P_d = f < 1$.

tapes, which contain balance sheets, income statements, and monthly stock prices for the 150 largest bank holding companies (BHC's). Although this sample is not representative of the entire population of all banks or bank holding companies, which comprises many smaller and often privately held organizations, the BHC's in this sample hold about 40 percent of all bank assets and thus are of interest in their own right. Data on the interest cost of large, uninsured CD's are from the Bank Consolidated Report of Conditions and Income (the Bank Call Report). Data on state branching, multibank holding company expansion, and interstate entry laws are from Amel and Keane (1987) and various Federal Reserve Annual Statistical Digests.

A. Measuring Market Power

As discussed above, q is used to measure the degree of market power in banking. The measure q is defined as the market value of assets (calculated as the sum of the market value of common equity—price per share times number of shares—and the book value of liabilities) divided by the book value of assets.⁸ The assumption is that the capitalized value of the bank charter will be reflected in the market value of equity (and thus the market value of assets as defined above), but not the book value of equity or assets.⁹ Thus, banks with larger charter val-

ues due to market power in asset and/or deposit markets should have greater market-to-book asset ratios. Note that the ability to issue deposits at below-market rates is an asset, the value of which will be reflected in the market value of the bank's equity and thus the market value of assets as I define them.

Several difficulties arise in using q as a measure of a bank's market power. First, the book value of assets represents the historical costs of assets acquired and sold over time, not the current costs of the assets. Thus, *ex post* market-to-book ratios that differ from 1 may reflect different asset return realizations rather than the degree of market power, which would be reflected in the *ex ante* q . Thus, the theoretically appropriate *ex ante* q is measured with error when using the *ex post* q . Another reason that q may not accurately reflect the degree of market power is that the value of potentially underpriced deposit insurance could be capitalized into a bank's market value.¹⁰ Several methods are used in the empirical analysis to control for these possibilities.

First, and most importantly, simultaneous equations techniques are used. In the first stage, an instrument is created for q , and then the predicted ratio is used as an explanatory variable in second-stage equations of bank risk taking. Thus, the empirical model allows for both the possible endogeneity of q and the fact that q is measured with error.

Second, a sample of banking organizations is selected to have similar histories. To do this, the sample is restricted to banking organizations for which data are continuously available from 1970 through 1986. Moreover, two variables are included to control for different asset histories: a dummy variable for banks that were on the Compustat tapes in 1964 and survived to be included in the sample (about 38 percent of

⁸Banks use an accounting convention in which loan loss reserves are counted as book capital. However, since loan loss reserves are often taken only when asset losses either have already occurred or when they are anticipated, they in fact often do not represent book capital since the addition to capital usually would be offset at least fully by asset losses if they were realized on the books. Thus, in this paper, loan loss reserves are not counted as book capital. (This procedure follows generally accepted accounting procedures as followed by bank holding companies in their 10-K and 10-Q reports filed with the Securities and Exchange Commission.)

⁹Although a firm that acquired a bank originally endowed with a valuable charter would record that charter's value (i.e., the excess of the purchase price over the book value) as an intangible asset, intangible assets are not counted as primary book equity, the measure of book equity used in this paper.

¹⁰As Smirlock and Gilligan (1984) argue, a q greater than 1 also could reflect the capitalized value of a firm-specific efficiency enhancing factor of production. However, a firm possessing such a factor would have the same incentives to protect its value as would a firm possessing market power.

the sample) and each bank's asset growth rate since 1970.

B. Model Structure

The empirical model consists of two sets of equations. In the first set of regressions, q is regressed on dummy variables that equal 1 during a given period if there was a liberalization in the laws governing state branching, multibank holding company, and interstate expansion, respectively, during any previous period. These regressions also contain a set of control variables and other proxies for market power such as the ratio of demand deposits to total deposits, the ratio of foreign deposits to total deposits, and the ratio of loans to assets. (The branching variables are constructed based on data in Amel and Keane [1987] and various Federal Reserve Statistical Abstracts.) The balance sheet data are from Compustat and refer to the consolidated holding company.

The hypotheses are that unanticipated liberalization of legal entry barriers should erode banks' market power and thus negatively affect q and that greater deposit funding and loan making also might be related to market power and thus positively affect q .

In a second set of regressions, both actual q and an instrument for q are explanatory variables in equations that attempt to explain bank risk taking. The hypotheses are that the banks with greater market power should have larger capital-to-asset ratios and lower risks of default.

The system to be estimated can be represented as

$$(14) \quad q_{it} = X_{1it}\beta_1 + \epsilon_{1it}$$

$$(15) \quad \text{risk}_{it} = X_{2t}\beta_2 + q_{it}\beta_3 + \epsilon_{2it}$$

where

X_{1it} is a vector of branching, financial, control, and other variables for bank i at time t ;

X_{2t} is a vector of financial and other control variables at time t ;

q_{it} is the bank's market-to-book asset ratio at time t [in the two-stage least-squares estimates, equation (14) is used to construct an instrument for q_{it}];

risk_{it} is a measure of bank default risk (two measures are used: the market-value capital-to-asset ratio [the market value of common equity divided by the market value of equity plus the book value of liabilities] and the interest cost on large CD's);

β_1 and β_2 are vectors of coefficients;

β_3 is the effect of q on risk; and

ϵ_{1it} and ϵ_{2it} are random error terms.

C. Empirical Results

Table 1 contains descriptive summary statistics for the bank holding companies in the sample using fourth-quarter (year-end) data for the 1970–86 period.¹¹ It is interesting to note that market-to-book asset ratios range from 0.95 to 1.18 and that market-value capital-to-asset ratios range from 0.0075 to 0.21. Although it is not shown in the table, about 32 percent of the bank holding companies were in states that liberalized branching laws, 44 percent were in states that liberalized multibank holding company expansion laws, and 78 percent were in states that liberalized interstate entry laws by 1986.

D. Market-to-Book Asset Ratios

Table 2 contains the results of the estimates of equation (14) over the 1971–86 period using fourth-quarter data. Estimates of the effects of four different types of variables are reported: branching variables, control variables, balance sheet variables, and financial variables. The branching variables are dummies reflecting the liberalization during a previous period of branching,

¹¹Some of the data items were missing on the Compustat tapes for particular bank holding companies in particular quarters. Rather than exclude the entire observation, missing values were forecast on the basis of quadratic time-trend OLS regressions estimated separately for each bank. For variables known a priori to be nonnegative, forecast values were constrained to be nonnegative.

TABLE 1—SUMMARY STATISTICS FOR 85 LARGE BANK HOLDING COMPANIES
(POOLED 1970–86, FOURTH-QUARTER DATA)

Characteristic	Mean	Minimum	Maximum
Liberalization of Branching Law (Dummy Variable)	0.19	0	1
Liberalization of Multibank Holding Co. Law (Dummy Variable)	0.26	0	1
Liberalization of Interstate Entry Law (Dummy Variable)	0.12	0	1
Book Value of Assets (Net of Loan Loss Reserves in \$ Millions)	\$10,587	\$278	\$195,147
Market-to-Book Asset Ratio, q	1.00	0.95	1.18
Multinational Regulatory Status (Dummy Variable)	0.19	0	1
Foreign Deposits/Total Deposits	14 percent	0 percent	89 percent
(Cash + Treasury Securities)/Total Assets	24 percent	7.8 percent	50.1 percent
Annual Assets Growth Rate Since 1970	12 percent	-1 percent	63 percent
Demand Deposits/Total Deposits	35.9 percent	0 percent	72.3 percent
Market-Value Capital-to-Asset Ratio	0.056	0.0075	0.21
Book-Value Capital-to-Asset Ratio	0.055	0.010	0.14
New York Composite Index	70.8	35.4	142.1
3-Month Treasury-Bill Rate	7.64 percent	4.02 percent	15.66 percent
20-Year Treasury-Bond Rate	9.03 percent	5.96 percent	13.73 percent
Average Maturity of CD's (Months)	6.45	1.19	19.81
Interest Cost of CD's	0.085	0.050	0.13

multibank holding company, and interstate expansion laws in the state in which the bank is located. The financial variables are the New York Composite Index, the three-month Treasury-bill rate, and the 20-year Treasury-bond rate. They are included to control for the effects of general interest rate and stock market trends on the market-to-book ratio that would not be related to changes in market power.

Other variables are included as proxies for market power. While some of them might be endogenous, the coefficient estimates of branching variables are not sensitive to their inclusion. The estimated coefficients generally conform with a priori expectations. Liberalization of branching or multibank holding company expansion laws in a previous period is associated with a statistically significantly (at the 1-percent level) lower market-to-book asset ratio. This suggests that both branching and multibank holding company expansion restrictions do provide banks a degree of protection from competition. This finding is consistent with a study by Mark J. Flannery (1984) that finds that unit banks in unit banking states

earn monopoly profits approximately 20 percent above those reported by similar banks in branching states, as well as other studies that have found that branching and multibank holding company expansion restrictions lead to higher loan rates and lower deposit rates.¹² (See Allen N. Berger and Timothy H. Hannan [1987] for recent evidence that restricted branching leads to lower deposit rates.) However, no significant effect of liberalization of interstate entry restrictions is found. This may reflect the fact that these laws generally allow entry only by acquisition (which would increase market prices) and do not allow de novo entry, which would directly increase competition and thus diminish market prices.

¹²There is an extensive literature on the effects of branching restrictions on competition. Generally, these studies show that branching restrictions are associated with reduced competition. For example, Donald T. Savage and Stephen A. Rhoades (1979) found banks in statewide branching states paid higher rates on deposits. Also, a survey by George J. Benston (1973) finds general gains in service for the banking public. See Savage and Elinor H. Solomon (1980) for a discussion of the literature.

TABLE 2—POOLED TIME-SERIES CROSS-SECTION REGRESSION FOR 85 LARGE BANK HOLDING COMPANIES RELATING THE MARKET-TO-BOOK ASSET RATIO TO VARIOUS DETERMINANTS OF MARKET POWER 1971–86, FOURTH-QUARTER DATA (STANDARD ERRORS IN PARENTHESES)

R ²	0.42
Number of Observations	1360
Dependent Variable Mean (Market-to-Book Asset Ratio, <i>q</i>)	1.00
Intercept	−0.94*** (0.011)
Branching Variables	
Liberalization of Branching Law	−0.0046*** (0.0017)
Liberalization of Multibank Holding Co. Law	−0.0074*** (0.0015)
Liberalization of Interstate Entry Law	0.00050 (0.0024)
Control Variables	
Dummy for Being on Compustat in 1964	0.0054*** (0.0013)
Dummy for Multinational Status	−0.0053** (0.0023)
Balance Sheet Variables	
Book-Value Asset Growth Since 1970	0.21*** (0.010)
Demand Deposits/Total Deposits (×100)	0.00056*** (0.000067)
Loans/Total Assets	0.0098 (0.012)
Foreign Deposits/Total Deposits (×100)	0.00019*** (0.000045)
Book Value of Assets	−4.09E ^{−9} (4.49E ^{−8})
(Cash and Treasury Securities)/Total Assets	−0.000027 (0.00013)
Financial Variables	
N.Y. Composite Index	0.0038*** (0.000030)
3-Month Treasury-Bill Rate	−0.00078** (0.00032)
20-Year Treasury-Bond Rate	−0.0019*** (0.00047)

*Significant at the 10-percent level.

**Significant at the 5-percent level.

***Significant at the 1-percent level.

The balance sheet variables generally have signs consistent with the notion that market power arises in deposit and loan markets, although statistically significant effects are found only for deposit markets. The fraction of demand and the fraction of foreign deposits in total deposits are positively and significantly related to the market-to-book asset ratio. The point estimate of the effect of the fraction of loans in total assets on the

market-to-book ratio is positive, and the estimate of the effect of the ratio of cash and treasury securities to total assets is negative, although neither variable is significant. Banks with more rapid asset growth in the past have significantly higher market-to-book ratios, perhaps because more rapid growth is associated with lack of competition or success due to other factors. Asset size per se, however, is not significant.

The control variable, whether a bank was on Compustat in 1964, is positively and significantly related to the market-to-book ratio, perhaps indicating that older, prominent bank holding companies that survived to be included in the sample have higher market values than bank holding companies that entered the sample later. The variable for multinational status is negatively related to the market-to-book ratio, which might be due to the very competitive international environment in which these 16 money center banks (as defined by the Federal Reserve) operate.

Finally, the effects of the financial variables are much as one might expect. Stock market values are positively related to the market-to-book ratio, and interest rates are negatively related.

Overall, the high correspondence between the expected effects of the variables and their estimated effects suggests that the market-to-book ratio is in fact a proxy for market power. Next, I test whether this proxy for market power is negatively related to bank risk taking.

E. Bank Risk

Below, the effects of the market-to-book asset ratio on two measures of bank default risk are examined. As discussed above, a key hypothesis is that the decline in banks' market power, as proxied by their market-to-book ratios, was a primary cause of the decline in banks' capital-to-asset ratios. Moreover, the theory implies that the cross-sectional variation in bank capital ratios would be influenced by variations in market power—banks with greater market power should have higher capital ratios.

F. Market-Value Capital-to-Asset Ratios

Table 3 presents coefficient estimates of equation (15), in which the market-value capital-to-asset ratio (i.e., the market value of capital divided by the market value of assets, defined as the market value of equity divided by the market value of equity plus the book value of liabilities) is regressed on the market-to-book asset ratio, q , holding

stock market and interest rate trends constant. It is necessary to hold stock market and interest rate trends constant since these trends potentially could influence both the dependent and independent variables, thus leading to spurious correlation. In addition, dummies for being on Compustat in 1964 and for multinational status are included.

In the first column of Table 3, ordinary least squares (OLS) estimates from a pooled time-series cross-section regression are reported. The OLS results suggest a strong, positive, and statistically significant relationship between the proxy for market power, the market-to-book asset ratio, and the market-value capital-to-asset ratio. Thus, as predicted, banks with more market power appear to hold more capital relative to assets. Moreover, the estimated magnitude of the effect is large, with a 10 percentage point increment in the market-to-book asset ratio leading to a 0.09 increase in the market-value capital-to-asset ratio, and is not sensitive to whether the variation in the market-to-book ratio is due to changes over time or differences across banks.¹³

There are, however, several reasons why the OLS estimates should be viewed with caution. First, endogeneity between q and bank risk is possible. For example, a bank

¹³To assess how sensitive the results were to pooling the cross sections over time, separate cross section regressions were run for each year from 1971 to 1986. Each of the OLS point estimates of the effect of the market-to-book asset ratio on the market-value capital-to-asset ratio were significantly different from zero at the 1-percent level and ranged from 0.62 to 1.19, approximately the same magnitude as the pooled time-series cross section results reported in Table 3.

Since different banks can have different responses to the market index (as proxied by the New York Composite), I also estimated an unconstrained version of equation (15) with separate intercepts and separate slope coefficients for the New York Composite Index for all 85 bank holding companies. However, the estimate of the effect of the market-to-book ratio was statistically significant and about the same magnitude as in the constrained model estimates reported in Table 3. Thus, the results appear robust regarding the source of variation in the market-to-book capital ratio—both the time-series and cross-sectional variation in banks' market-value capital-to-asset ratios are positively associated with time-series and cross-sectional variation in banks' market-to-book capital ratios.

TABLE 3—POOLED TIME-SERIES CROSS SECTION REGRESSION RELATING THE MARKET-VALUE CAPITAL-TO-ASSET RATIO TO THE MARKET-TO-BOOK ASSET RATIO AS A PROXY FOR MARKET POWER, FOURTH-QUARTER DATA, 1971–86
(STANDARD ERRORS IN PARENTHESES)

	OLS	TSLS ^a	TSLS ^b
\bar{R}^2	0.83	0.66	0.15
Number of Observations	1360	1360	1360
Dependent Variable Mean (Market-Value Capital-to-Asset Ratio)	0.054	0.054	0.054
Intercept	−0.85*** (0.014)	−0.70** (0.026)	−1.00*** (0.29)
N.Y. Composite Index	−0.0000044 (0.000013)	0.000032** (0.000015)	−0.000052 (0.000029)
3-Month Treasury-Bill Rate	−0.00042** (0.00018)	−0.00043** (0.00018)	−0.00042 (0.00051)
20-Year Treasury-Note Rate	−0.00023 (0.00025)	−0.00091*** (0.00028)	0.00065 (0.0015)
Dummy for Being on Compustat in 1965	0.0018** (0.00073)	0.0021 (0.00076)	−0.054 (0.046)
Dummy for Multinational Status	−0.018*** (0.00094)	−0.018*** (0.00098)	−0.032** (0.015)
Market-to-Book Asset Ratio (q)	0.91*** (0.013)	0.77*** (0.025)	1.09*** (0.27)

*Significant at the 10-percent level.

**Significant at the 5-percent level.

***Significant at the 1-percent level.

^aIncludes as instruments all variables on right-hand side of regression reported in Table 2.

^bIncludes as instruments only the branching, multibank holding company, and interstate expansion dummies, the financial variables, the on-Compustat dummy, and the dummy for multinational status.

with greater default risk could have a greater market-to-book asset ratio if deposit insurance were underpriced and its value were capitalized in a bank's market (but not book) value. Second, the market-to-book asset ratio measures market power with error due to *ex post* asset return realizations that are different from *ex ante* expectations. Third, q and the market-value capital-to-asset ratio might be spuriously correlated due to the presence of the market value of the bank's equity on both sides of the equation. Although the equation is not an identity, to the extent the ratio of the book value of liabilities to the book value of assets were approximately constant or much less variable than the ratio of the market value of equity to the book value assets, the estimated OLS coefficient on q would be biased toward 1. For these reasons, a simultaneous equations model is employed, and two-stage least squares (TSLS) estimates also are displayed in Table 3. By employing

TSLS techniques, exogenous variation in q is related to actual variation in market-value capital-to-asset ratios thus avoiding the potential endogeneity and measurement error problems associated with q . TSLS techniques also solve the problem of potential spurious correlation since the actual market value of common equity is not a right-hand-side variable. (The method of estimation employed produces standard errors corrected for the two-step nature of the procedure.)

In the first-column TSLS estimates, the instruments used to predict the market-to-book capital ratio include all of the explanatory variables in equation (14). In the second-column TSLS estimates, only the branching, multibank holding company and interstate expansion dummies, the financial variables, the on-Compustat dummy, and the multinational dummy were included as instruments, variables believed to be exogenous. Thus, in these second-column esti-

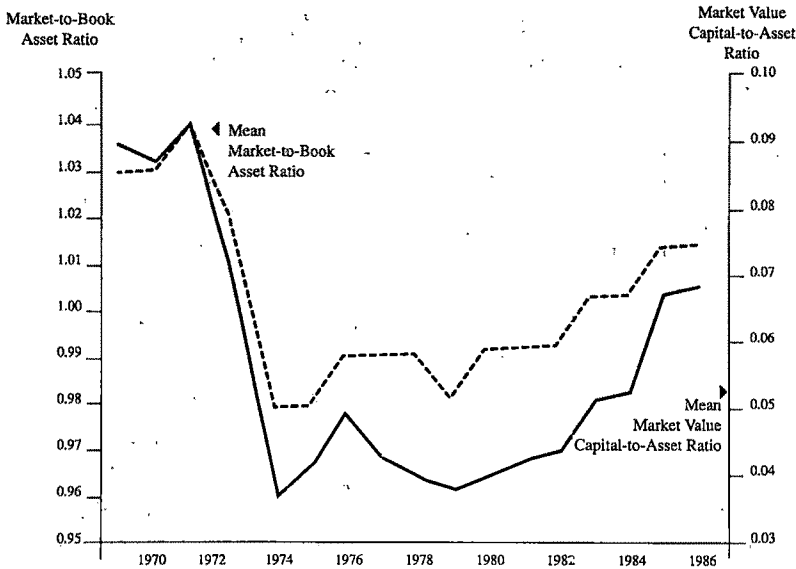


CHART 3. MEAN MARKET-TO-BOOK ASSET RATIO AND MARKET-VALUE CAPITAL-TO-ASSET RATIO

mates, the exogenous variation in q is due mainly to exogenous variation in the branching variables. However, the TSLS estimates of the effect of the market-to-book capital ratio on the market-value capital-to-asset ratio are very similar to the OLS estimate, especially the second TSLS estimate. This may be because the biases due to measurement error and endogeneity due to the capitalization of underpriced deposit insurance are offsetting. Thus, the finding that banks with more market power hold more capital relative to assets is robust with respect to the estimation method and specification of the model.

Moreover, the decline over time in banks' market-value capital-to-asset ratios is strongly associated with the decline in their market-to-book asset ratios. Chart 3 shows that the mean of banks' market-value capital-to-asset ratios follows the mean of their market-to-book ratios over time. According to the theory developed in Furlong and Keeley (1987, 1989), banks with more capital have less incentive to increase asset risk. Thus, as long as the stringency of asset risk regulation is not less at banks with stronger capital positions, such banks should have

lower default rates and thus represent lower risk exposures to the FDIC.¹⁴ In the next section, I present a test of this hypothesis, proxying risk exposure with the interest cost on large, uninsured CD's.

G. Interest Cost of Large CD's

Ideally, one would like to relate market power in banking directly to the risk exposure of the FDIC and a bank's uninsured depositors. While the FDIC's risk exposure is not directly observable,¹⁵ one can obtain a

¹⁴As Furlong and Keeley (1987a, 1989) show, the incentive to increase asset risk declines as leverage declines because the second derivative of the value of the insurance put option with respect to asset risk with respect to leverage is positive. This result holds in a multistate state preference or a continuous option pricing framework, but it does not hold for the simple two-state model presented in this paper.

¹⁵As Marcus and Shaked (1984) and Ronn and Verma (1986) have shown, it is possible to estimate the FDIC's risk exposure by using an options pricing model that relates the observed risk of bank equity (with deposit insurance) to the unobserved risk of bank assets (absent deposit insurance). However, such estimates require assumptions about bank closure policy, which bank and bank holding company liabilities are

measure of the risk premium on a bank's uninsured deposits, assuming that uninsured depositors behave as if they are not implicitly fully insured. The hypothesis is that the rate on large (over \$100,000) certificates of deposit (CD's) contain a risk premium related to the bank's default risk, which should be negatively related to a bank's market power as reflected in its market-to-book asset ratio, q .

Although there is a debate regarding whether, in fact, large CD's are implicitly fully insured (and thus whether they contain a risk premium), recent empirical evidence in Timothy H. Hannan and Gerald Hanweck (1988) and Christopher James (1987) strongly suggests that they do contain a risk premium. Moreover, if large CD's are sold in a national (or international) market, it is hard to imagine other factors that could explain the rate differences among banks on CD's of identical maturity.

The interest cost of large CD's is estimated from information contained in the Bank Consolidated Report of Condition and Income (the Bank Call Report). The average rate on large CD's is estimated by dividing the total interest paid (by all of the banks in the bank holding company) by the average dollar value outstanding during the year. Because of difficulties in constructing consolidated interest costs and amounts outstanding for all of the 85 bank holding companies in the previous sample, the sample had to be restricted to 77 bank holding companies for which complete information could be obtained.

Following James (1987), a weighted average maturity of CD's also is constructed to control for differences in the maturities of CD's outstanding. Since the data needed to construct this variable have only been collected since 1984, only the 1984-86 data are

used. Following James, I also control for time-series variation in the level of interest rates by including the average yield on three-month Treasury bills in the CD rate regressions.

The regression results are reported in Table 4. In the first column, estimates of the effect of the market-value capital-to-asset ratio on the CD rates are presented. As mentioned above, banks with more market power, as reflected in higher market-to-book asset ratios, hold more capital relative to assets, in theory, in order to protect their valuable charters. If this theory is correct, banks with more capital relative to assets should have lower default probabilities and thus should have lower CD rates. The results in the first column of Table 4 confirm this hypothesis: banks with greater market-value capital-to-asset ratios pay lower CD rates. In fact, a 1 percentage point increase in a bank's capital-to-asset ratio would lower its CD rate by 14 basis points.¹⁶ This result is somewhat larger but comparable to the 8-basis-point effect Hannan and Hanweck (1988) found using the book-value capital-to-asset ratios in a somewhat different specification.

In the second and third columns of Table 4, estimates of the effects of the market-to-book asset ratio on the CD rate are presented. As in the previous equations reported in Table 3, both OLS and TSLS results are presented. The TSLS results are obtained by using all of the right-hand-side variables in equation (14) (less one interest rate trend variable, due to degrees of freedom limitations) as instruments. The results confirm the hypothesis that banks with more market power have lower default risk.

As expected, the coefficient on the market-to-book asset ratio is negative and significantly different from zero at the 1-percent level. Moreover, the effect, while not large, is economically meaningful (each 1 percentage point increment in the market-to-book asset ratio reduces the average CD

insured, and a number of other factors (such as the market value of preferred stock, whether bank holding company assets will be used to support a failing bank, etc.). Because of uncertainty regarding just what assumptions might be appropriate and possible changes in a number of these factors over time, I focus on CD rates, a more directly observable proxy for bank default risk.

¹⁶This result is also consistent with the hypothesis that banks with larger capital-to-asset ratios have less incentive to increase asset risk.

TABLE 4—POOLED TIME-SERIES CROSS SECTION REGRESSION RELATING INTEREST COST OF LARGE CD'S TO THE MARKET-TO-BOOK ASSET RATIO AS A PROXY FOR MARKET POWER AND THE MARKET-VALUE CAPITAL-TO-ASSET RATIO, FOURTH-QUARTER DATA, 1984–86 (77 BANK HOLDING COMPANIES)
(STANDARD ERRORS IN PARENTHESES)

	OLS	OLS	TSLs
\bar{R}^2	0.43	0.43	0.39
Number of Observations	231	231	231
Dependent Variable Mean (Interest Cost of CD's Divided by CD's Outstanding)	0.085	0.085	0.085
Intercept	0.029*** (0.0064)	0.020*** (0.0043)	0.19*** (0.070)
3-Month Treasury-Bill Rate	0.83*** (0.074)	0.82*** (0.074)	0.81*** (0.084)
Average Maturity of CD's	0.0011*** (0.00025)	0.0011*** (0.00025)	0.00073 (0.00055)
Market-to-Book Asset Ratio (q)		-0.18*** (0.042)	-0.16*** (0.066)
Market-Value Capital-to-Asset Ratio	-0.14*** (0.034)		

*Significant at the 10-percent level.

**Significant at the 5-percent level.

***Significant at the 1-percent level.

cost by 16–18 basis points). It is important to recognize that this estimated effect arises from both the time-series as well as the cross-sectional variation in CD costs. This result is consistent with the significant positive effect of the market-to-book ratio on the market-value capital-to-asset ratio reported in Table 3 and the significant negative relationship between the market-value capital-to-asset ratio and the CD risk premium reported in column one of Table 4.

III. Summary and Conclusion

This paper addresses two major empirical puzzles. Why has the deposit insurance system worked as well as it has over much of its history even though it provides a moral hazard for excessive risk taking, and why is the cross-sectional distribution of bank risk taking nonuniform? The hypothesis is that various anticompetitive restrictions endowed banks with market power and made banking charters valuable. The potential loss of a charter in the event of bankruptcy created, in effect, a regulatory bankruptcy cost, which counterbalanced the incentive

for excessive risk taking due to fixed-rate deposit insurance.

The empirical results are consistent with this hypothesis. Banks with more market power, as reflected in larger market-to-book asset ratios, hold more capital relative to assets (on a market-value basis) and they have a lower default risk as reflected in lower risk premiums on large, uninsured CD's. Thus, at least some of the increase in bank and thrift failures and payouts from the deposit insurance funds may be due to a general decline in the value of bank charters associated with increased competition within the banking and financial service industry.

In the past, the perverse incentives created by the deposit insurance system were countervailed by the potential loss of a valuable charter that induced banks to limit their own risk taking. This does not mean that it is desirable or even possible to return to a system of anticompetitive restrictions in order to reduce banking risk. But it does mean that the deposit insurance system must be reformed to reduce the rewards it provides for excessive risk taking.

REFERENCES

- Amel, Dean F., and Keane, Daniel G., "State Laws Affecting Commercial Bank Branching, Multibank Holding Companies, and Interstate Banking," Working Paper, Board of Governors of the Federal Reserve System, 1987.
- Benston, George J., "The Optimal Banking Structure," *Journal of Bank Research*, Winter 1973, 3, 220-37.
- Berger, Allen N. and Hannan, Timothy H., "The Price Concentration Relationship in Banking," Working Paper No. 100, Board of Governors of the Federal Reserve System, June 1987.
- Dothan, Uri and Williams, Joseph, "Banks, Bankruptcy, and Public Regulation," *Journal of Banking and Finance*, March 1980, 4, 65-87.
- Flannery, Mark J., "The Social Costs of Unit Banking Restrictions," *Journal of Monetary Economics*, March 1984, 13, 237-49.
- Furlong, Frederick T. and Keeley, Michael C., "Bank Capital Regulation and Asset Risk," *Economic Review*, Federal Reserve Bank of San Francisco, Spring 1987, 20-40.
- _____, "Bank Capital Regulation and Risk Taking: A Note," *Journal of Banking and Finance*, November 1989, 13, 883-91.
- Hannan, Timothy H. and Hanweck, Gerald, "Bank Insolvency Risk and the Market for Large CDs," *Journal of Money, Credit, and Banking*, May 1988, 20, 438-46.
- James, Christopher, "Off-Balance-Sheet Banking," *Economic Review*, Federal Reserve Bank of San Francisco, Fall 1987, 21-36.
- Kareken, John H. and Wallace, Neil, "Deposit Insurance and Bank Regulation: A Partial Equilibrium Exposition," *Journal of Business*, July 1978, 51, 413-38.
- Keeley, Michael C., (1985a) "The Regulation of Bank Entry," *Economic Review*, Federal Reserve Bank of San Francisco, Summer 1985, 5-13.
- _____, (1985b) "Bank Entry and Deregulation," *Weekly Letter*, Federal Reserve Bank of San Francisco, August 25, 1985.
- _____, "Bank Capital Regulation in the 1980s: Effective or Ineffective," *Economic Review*, Federal Reserve Bank of San Francisco, Winter 1988, 1-20.
- Kling, Arnold, "The Banking Crisis from a Macroeconomic Perspective," Working Paper, Board of Governors of the Federal Reserve System, 1986.
- Lindenberg, Eric and Ross, Stephen, "Tobin's q Ratio and Industrial Organization," *Journal of Business*, January 1981, 54, 1-32.
- Marcus, Alan J., "Deregulation and Bank Financial Policy," *Journal of Banking and Finance*, December 1984, 8, 557-65.
- _____, and Shaked, Israel, "The Valuation of FDIC Deposit Insurance Using Option-Pricing Estimates," *The Journal of Money, Credit, and Banking*, November 1984, Part 1, 16, 446-60.
- Merton, Robert C., "An Analytic Derivation of the Cost of Deposit Insurance Loan Guarantees," *Journal of Banking and Finance*, June 1977, 1, 3-11.
- Peltzman, Sam, "Entry into Commercial Banking," *Journal of Law and Economics*, October 1965, 8, 11-50.
- Pennacchi, George, "A Reexamination of the Over- (or Under-) Pricing of Deposit Insurance," *Journal of Money, Credit, and Banking*, August 1987, 19, 340-60.
- Ronn, Ehud and Verma, Avinash K., "Pricing Risk-Adjusted Deposit Insurance: An Option-Based Model," *Journal of Finance*, September 1986, 41, 871-94.
- Salinger, Michael, "Tobin's q , Unionization, and the Concentration-Profits Relationship," *Rand Journal of Economics*, Summer 1984, 15, 159-70.
- Savage, Donald T. and Rhoades, Stephen A., "The Effect of Branch Banking on Pricing, Profits, and Efficiency of Unit Banks," in *Proceedings of a Conference on Bank Structure and Competition*, Federal Reserve Bank of Chicago, 1979.
- _____, and Solomon, Elinor H., "Branch Banking: The Competitive Issues," *Journal of Bank Research*, Summer 1980, 11, 110-21.
- Sharpe, William F., "Bank Capital Adequacy, Deposit Insurance, and Security Values," *Journal of Financial and Quantitative*

- Analysis Proceedings*, November 1978, 13, 701-18.
- Smirlock, Michael, Gilligan, Thomas and Marshall, William, "Tobin's q and the Structure Performance Relationship," *American Economic Review*, December 1984, 74, 1051-60.
- Stigler, George, "A Theory of Oligopoly," *Journal of Political Economy*, February 1964, 72, 44-61.
-

Overdrafts and the Demand for Money

By AVNER BAR-ILAN*

This paper presents a stochastic analysis of the demand for interest-bearing money, such as NOW accounts, when overdrafting is allowed at some penalty rate. It is shown that the short-run interest elasticity of money demand is probably large (in absolute value) and negative, but in the long run this elasticity is much smaller or even positive. It is also argued that current definitions of the monetary aggregates, which exclude unused credit, may spuriously generate instability of money demand. An alternative definition of money stock is suggested, and seems to be conceptually more satisfying. (JEL 311)

The proposition that unused credit should count as money goes back at least to Keynes, who wrote in 1930:

There exists in unused overdraft facilities a form of Bank-Money of growing importance, of which we have no statistical record... the Cash Facilities, which are truly cash for the purposes of the Theory of the Value of Money, by no means correspond to the Bank Deposits which are published. The latter... take no account of something which is a Cash Facility, in the fullest sense of the term, namely, unused overdraft facilities.

[Keynes, 1930 pp. 42-43]

Although many economists, both before and after Keynes, have expressed similar views,¹ no explicit derivation of the demand for money with overdrafts has been carried out. This paper is an attempt to remedy this omission. In addition, it generalizes previous models by allowing some components of the money supply, such as NOW accounts, to bear interest. This generates rich dynamics of the response of the money stock to changes in interest rates.

*Department of Economics, Tel-Aviv University, 69978, Tel-Aviv, Israel. My thanks to Alex Cukierman, Benjamin Eden, M. Jane Flanders, Meir Kohn, Alex Zanello, and two referees for their very useful comments. Financial support from the Foerder Institute for Economic Research is gratefully acknowledged.

¹Two of the numerous examples are Lavington (1921) and Laffer (1970).

According to the transactions theory of money demand,² the optimal rule of money holding is a trigger-target rule; that is, the money stock is adjusted to the target only when it falls below the trigger.³ However, an assumption that is common to virtually all papers in the field is to constrain the trigger to an exogenous value, which is usually zero.⁴ Optimization is then carried out on the target level only. In the solution presented here, both the target and the trigger are chosen optimally. This is accomplished by using impulse control, a relatively new technique of optimal control.⁵

A new definition of the money supply, which is closely related to the one offered by Keynes, is suggested. By including outstanding credit, the new definition seems conceptually more appropriate as a measure of the quantity of the medium of exchange. Moreover, it means that current definitions of the money stock, which assign zero weight to outstanding credit, may impart a consis-

²The transactions theory of money demand originated in a deterministic framework due to Baumol (1952) and Tobin (1956) and a stochastic version by Miller and Orr (1966). Some of the recent examples that extended these original works are Milbourne, Buckholtz, and Wasan (1983) and Romer (1986, 1987). The most general solution, and the one that is closest to this paper, is that of Frenkel and Jovanovic (1980).

³The reason for the optimality of this rule is a fixed transaction cost in converting bonds to money.

⁴Equating the trigger level to zero is implicitly equivalent to excluding the possibility of overdrafting.

⁵Foundations of impulse control can be found in Bensoussan and Lions (1982).

tent bias to the measures of money. This can shed some light on the difficulties in applying the transactions theory in order to measure money (for example, the "missing money" puzzle of Goldfeld [1976]) or the excess volatility of the velocity of money.

The structure of the paper is as follows. Section I demonstrates some of the important insights in a simple, deterministic model. Section II presents the problem of optimal money holding with a stochastic disbursement process, while Section III describes the solution. Section IV discusses some of the implications of this solution, and Section V elaborates on the consequences for the definition of money. Concluding comments are offered in Section VI.

I. A Simple Deterministic Model

Some of the basic insights of the general case can be demonstrated in a deterministic framework with no discounting. Consider the portfolio choice of an individual or a business firm. There are two assets: money, the medium of exchange, and another asset, called "bonds," which cannot be used as a means of payment. Hence, people must hold money to complete their transactions even though they implicitly pay a liquidity premium for doing so, since money yields less interest than bonds. The amount of money held is determined by minimizing the expected costs associated with holding money. These costs take the following form:

(i) The cost of holding a money balance m is denoted by $I(m)$. When $m > 0$, the cost is the forgone interest on bonds relative to money; when $m < 0$, the cost is a shortage cost that is the excess interest paid on overdrafts relative to bonds, or any other penalty paid on negative account balances. Assume that both the holding and penalty costs are linear to get

$$(1) \quad I(m) = \begin{cases} rm & \text{for } m \geq 0 \\ -pm & \text{for } m < 0 \end{cases}$$

where $r > 0$ is the difference between the interest rate on bonds and that on positive

money balances and $p > 0$ is the cost per dollar of holding negative money balances.

(ii) The cost of transfer of u dollars from bonds to money is denoted by $C(u)$. These costs might include two terms: a fixed cost K per transfer, which is independent of the transaction size, and a proportional brokerage fee c per dollar. This gives

$$(2) \quad C(u) = \begin{cases} K + cu & \text{for } u > 0 \\ 0 & \text{for } u = 0 \end{cases}$$

with $K, c > 0$.⁶

Consider now the straightforward extension of the Baumol-Tobin analysis of money demand by allowing overdrafting at the penalty rate p , as in equation (1). Assuming a constant rate of expenditures, the money stock bounces between a lower level μ and an upper level M with a sawtooth shape. At time $t = 0$, when the account balance is μ , an amount of g dollars is paid in bonds to be spent during the period (which is of length 1). At $t = 0$, the consumer makes the first of n bond sales each of size $(M - \mu)$. With no proportional cost [$c = 0$ in equation (2)] and no discounting, the transactions cost is

$$C = nK.$$

Assuming $\mu \leq 0$, the holding cost is

$$I = n \left(\frac{M}{g} \right) \left(\frac{M}{2} \right) r + n \left(\frac{\mu}{g} \right) \left(\frac{\mu}{2} \right) p.$$

Minimization of the total cost ($C + I$) with respect to M , μ , and n subject to the

⁶Implicit in equation (2) is the assumption that transfers from money to bonds ($u < 0$) are prohibited because of the very large cost of transfer in this direction. The assumption is made for computational reasons by allowing for only one trigger point, from bonds to money, and excludes the upper point that might trigger a transfer from money to bonds. The implications of this assumption are less important when the downward drift of the money stock is large relative to the standard deviation. See Frenkel and Jovanovic (1980 footnote 3).

constraint

$$n(M - \mu) = g$$

yields the following solution:

$$(3) \quad M = \left(\frac{2gKp}{r(p+r)} \right)^{1/2}$$

$$(4) \quad \mu = -\frac{r}{p}M.$$

Optimizing again, but assuming $\mu \geq 0$, gives the internal solution $M = \mu$, which maximizes total cost. The solution of minimum cost in this case is the corner solution $\mu = 0$, which is inferior to the solution given in equations (3) and (4). I conclude that the levels M and μ in equations (3) and (4) minimize the total cost.

What distinguishes the solution (3)–(4) from the Baumol-Tobin case is the use of overdrafts ($\mu < 0$) in the model, for any finite value of interest rates r and p . Only when the penalty rate p of using the credit is infinitely high (relative to the interest rate r) do equations (3) and (4) reduce to the well-known Baumol-Tobin result of $\mu = 0$ and $M = (2gK/r)^{1/2}$. Unless it is prohibitively expensive, individuals will utilize their available credit.

The intuition of this result, which holds also in the general case of stochastic disbursements with discounting, is the following. As long as the money balance is positive, the optimal policy is not to increase it but rather to let it fall at the rate of expenditures. This delaying of action reduces both transaction costs (since transactions are less frequent) and holding costs, since the average amount of money held is lower. Similarly, when the money stock is zero, it pays to wait, at least a little while, before converting bonds to money. In addition to saving on transactions cost, the penalty cost, proportional to the amount of credit used, is small for low levels of overdrafts.

The robust conclusion of the transactions theory of the demand for money is that

when credit is available to firms or individuals, even at a cost, they will frequently utilize it.⁷ It is interesting to compare this conclusion with that of Lucas and Stokey (1983). Their model, like mine, attempts to explain the use of both money and credit. In both models, money and credit are used to facilitate transactions, not as wealth. However, in Lucas and Stokey's model, money can purchase any good, while credit can be used to purchase only some goods ("credit goods"), but not others ("cash goods").

As a result, the two means of payment cannot be perfect substitutes in Lucas and Stokey's framework. In fact, the degree of substitutability between money and credit depends on two exogenously given factors: the size of the subset of credit goods and the substitutability in consumption of credit and cash goods (which is a property of individuals' preferences for these goods). Hence, the relative use of credit and money depends not only on the relevant interest rates (r and p in my notation), but also on two exogenously given factors. No matter how large the cost r of holding money is, the consumer cannot substitute credit for cash if he wishes to consume cash goods.

Here, on the other hand, money and credit are inherently very close substitutes, since both are perfectly acceptable means of payment (in the Lucas and Stokey terminology, all goods are "credit goods"). The utilization of credit and money depends only on the relative costs. An increase in the cost of holding money relative to credit (higher r/p ratio) results in substitution of credit for money, as given by equations (3) and (4). This sensitivity with respect to interest rates is preserved also in the general case, since the two means of payment provide a similar amount of liquidity.

Another issue that will be discussed later and can be demonstrated by using the simple model of this section is the appropriate definition of the money stock. There are at

⁷The case for using credit is even stronger with discounting. In this case it is more profitable to postpone the payment of the finite cost K , since the higher penalty cost is paid later. This constitutes an additional reason to reduce μ .

least three possible definitions:

$$(5) \quad E_1(m) = \frac{1}{2}(M + \mu)$$

$$(6) \quad E_2(m) = \frac{M^2}{2(M - \mu)}$$

$$(7) \quad E_3(m) = \frac{1}{2}(M - \mu).$$

$E_1(m)$ is the average account balance, which is positive only when $r < p$. $E_2(m)$ is the standard definition of money, the sum of positive balances in checking accounts, which assigns zero weight to overdrafts. Observing the liquidity of credit, whether used or not, $E_3(m)$ measures the amount of available means of payment at each point in time up to the level μ . The three definitions satisfy the inequality $E_3(m) \geq E_2(m) \geq E_1(m)$, whereas the equality holds in the Baumol-Tobin case. The claim that I make in Section V is that the more appropriate measure of the medium of exchange is given by the broader aggregate $E_3(m)$.

II. Formulation of the Stochastic Problem

At each point in time t , when the money stock is $m(t)$, the agent decides whether to convert bonds to money. Suppose he decides on such a transfer of size u_i dollars at time t_i . This transfer costs $C(u_i)$, defined in equation (2), and is carried out promptly to yield the money stock $m(t_i^+) = m(t_i) + u_i$.

The money stock on hand is also changed by the random net expenditure flow, according to the following stochastic differential equation:

$$(8) \quad dm(t) = -gdt + \sigma dw(t) + \sum_{i \geq 1} u_i \delta(t - t_i).$$

Positive values of mean disbursements g denote net cash outflow. The stochastic part of the expenditures is described by the Wiener process $w(t)$ with mean zero and

variance t .⁸ The last term in (8) denotes the discrete⁹ increases of size u_i of the money stock made at times t_i where $\delta(t)$ is Dirac's delta function.¹⁰

The optimizing agent chooses a sequence of financial transfers u_i made at t_i in order to minimize the expected discounted cost over an infinite horizon:

$$(9) \quad V(m) = \min_{\{u_i, t_i\}} E_0 \left[\int_0^\infty I(m(t)) e^{-\alpha t} dt + \sum_{i \geq 1} (K + cu_i) e^{-\alpha t_i} \right]$$

subject to the stochastic process described in equation (8). E_0 denotes the expectations operator given information known at time zero, and α is the interest rate on bonds. The cost of holding (positive or negative) money stock $m(t)$ is accumulated continuously at a rate $I(m(t))$ given in equation (1). The transfer cost $K + cu$ is incurred discretely.

⁸Frenkel and Jovanovic (1980) identify g as representing the transactions motive for holding money, while σ^2 stands for the precautionary motive. Miller and Orr (1966), on the other hand, interpret σ^2 loosely as a transactions term (p. 425). I think the latter interpretation is more appropriate, because there is no room for precautionary motives in this framework. Since the analysis is done in continuous time and the disbursement flow is finite, there is zero probability that money holding will overshoot the thresholds. In this case, even when σ^2 is large, the agent can control his money holding by choosing the right thresholds without worrying about holding money as a precaution against an unexpectedly low stock. In order to study the precautionary motive, the analysis should be done in discrete time, when there might be a finite probability of overshooting the trigger levels.

⁹Financial transactions will be made infrequently because of the fixed cost $K > 0$ that accompanies any transaction. In this case, a continuous transfer during any finite period of time results in an infinitely high cost.

¹⁰The delta function is defined by

$$\int_a^b f(x) \delta(x - c) dx = \begin{cases} f(c) & \text{if } a < c < b \\ 0 & \text{otherwise} \end{cases}$$

for any continuous function $f(x)$.

A common assumption made in studies of money demand is that the optimal transfer policy $\{u_i, t_i\}$ takes the form of simple trigger-target rules. The existence of such a rule for the problem (8)–(9) was established by Constantinides and Richard (1978). They proved that the optimal policy is of the (S, s) type studied in the inventories literature: when the money stock is below the trigger point s , a sale of bonds will be made such that the quantity of money will increase to the target level S ; otherwise no financial transaction will be made. I can now proceed to the evaluation of these trigger and target levels, denoted by μ and M , respectively, from now on.

III. Solution

The optimization problem described in the former section generalizes previous work on money demand in several ways. The most important is the consideration of overdrafts. This option has been implicitly excluded in other work; instead the trigger level μ was assumed to have a certain value (usually zero), and the solving procedure has been to find the target M , given the exogenous value of μ .

Allowing overdrafting at a finite penalty rate makes the trigger μ a control variable that is chosen optimally. Hence the solution to the optimization problem requires finding both μ and M . This can be accomplished by using an optimal control theory, the "impulse control," which analyzes optimal behavior in continuous time with fixed cost of taking an action.¹¹ This type of control characterizes the behavior of the cash manager in the presence of a fixed transactions cost. Sulem (1986) used this apparatus to solve the optimization problem (8)–(9) as follows. At each point in time t , when the money stock is at a level m , the cash manager can either sell bonds immediately to increase his money stock or postpone his transaction at least to time $t + \tau$. In the

latter case

$$(10) \quad V(m) \leq \int_t^{t+\tau} I(m(x)) e^{-\alpha(x-t)} dx + V(m(t+\tau)) e^{-\alpha\tau}.$$

The first term on the right-hand side of equation (10) is the cost of money holding between t and $t + \tau$. From Bellman's principle of optimality, the second term in (10) is the minimum expected cost from time $t + \tau$ on. Expanding equation (10) as a Taylor series around time t , where $m(t + \tau) = m - g\tau + \sigma dw(t)$ from equation (8) and letting $\tau \rightarrow 0$ yield the following differential equation:¹²

$$(11) \quad -\frac{1}{2}\sigma^2 \frac{d^2V}{dm^2} + g \frac{dV}{dm} + \alpha V \leq I.$$

When bonds are sold at time t , the money stock increases immediately to level $m + u$. Since the transaction size u is chosen optimally, then,

$$(12) \quad V(m) \leq K + \min_{u \geq 0} (cu + V(m + u)).$$

Since either equation (11) or (12) must hold as an equality, $V(m)$ is the solution of the following set of equations:

$$(13) \quad AV \leq I$$

$$(14) \quad V \leq BV$$

$$(15) \quad (AV - I)(V - BV) = 0$$

where

$$AV(m) \equiv -\frac{1}{2}\sigma^2 \frac{d^2V}{dm^2} + g \frac{dV}{dm} + \alpha V$$

$$BV(m) \equiv K + \min_{u \geq 0} (cu + V(m + u)).$$

¹¹Constantinides and Richard (1978) have used this theory to prove the optimality of (S, s) as the money rule for the problem (8)–(9).

¹²The properties of mean zero and variance t of the Wiener process $w(t)$ are also used in the derivation of equation (11). The function I is defined in equation (1). An equation similar to (11) is derived in Dixit (1989).

The system in (13)–(15), called a quasi-variational inequality in the impulse control literature, allows solving for the expected cost V as a function of the money stock m and for the trigger and target levels. The solution is stated in the following theorem.

THEOREM: *The optimal levels of the target M and the trigger μ are given by the following equations:*

$$(16) \quad M = (\alpha c + r)^{-1} \left[(\alpha c - p)\mu + \left(\frac{p + r}{\lambda_2} \right) (e^{\lambda_2 \mu} - 1) - \alpha K \right]$$

$$(17) \quad e^{-\lambda_1 M} = (\alpha c + r)^{-1} \left[(\alpha c - p)e^{-\lambda_1 \mu} + \left(\frac{p + r}{\lambda_1 - \lambda_2} \right) (\lambda_1 e^{(\lambda_2 - \lambda_1)\mu} - \lambda_2) \right]$$

where the parameters λ_1 and λ_2 are defined as

$$(18) \quad \lambda_1 = \sigma^{-2} \left[-(g^2 + 2\alpha\sigma^2)^{1/2} + g \right] \leq 0$$

$$(19) \quad \lambda_2 = \sigma^{-2} \left[(g^2 + 2\alpha\sigma^2)^{1/2} + g \right] \geq 0.$$

PROOF:

See Appendix 1.

IV. Analysis of the Solution

A. Discussion

The first property of equations (16) and (17) worth mentioning is that the solution for M and μ is homogeneous of degree one in the vector (g, σ, K) . Hence money demand is demand for real balances.

The framework for analyzing money demand described in the preceding two sections generalizes previous research in three respects.

(i) Overdrafting is allowed at some finite penalty rate, p .

(ii) The cost of holding money, r , is not necessarily equal to the discount rate, α .

(iii) The proportional cost of transferring bonds to money, c , is not restricted to zero.

One of the important properties of the solution (16)–(17), as demonstrated in Appendix 1, is $\mu \leq 0$. The prevailing assumption that constrains the trigger level to zero is therefore literally correct only when the cost of credit is infinitely high.¹³ Thus, $\mu \leq 0$ is an unambiguous prediction of the transactions theory of the demand for money, a prediction that holds in the general stochastic case and not only in the simple model of Section I; economic agents do use their credit provision when it is not excessively expensive.¹⁴

The assumption of the equality between r and α , which is also made very often in the literature, is a very limiting one. If the cost of holding money, r , is assumed to equal the bond market rate, α , it must be the case that the interest paid on money is zero. This might have been a good assumption when the interest paid on demand deposits was legally so constrained. In the wake of the deregulation of the banking industry, the assumption $r = \alpha$ restricts the analysis to the demand for currency, which is not what the transaction theory of money demand is about. Since a large fraction of the monetary aggregates bear positive interest rates, allowing for $r < \alpha$ is crucial for analyzing the demand for money.

In general, there is a dichotomy in the literature between models with fixed cost K only, as in the Baumol-Tobin or Miller and Orr (1966) models, and analyses of proportional cost with no fixed cost, as in Eppen and Fama (1969). However, the consequences of the inclusion of both sources of

¹³Numerical solutions for μ and M , some of them shown below (for example, Figs. 1 and 2), show that for a wide range of parameters the convergence of μ to zero when p increases is fairly slow.

¹⁴Empirical support of a widespread use of overdrafts and trade credit appears in Kanninen (1978), Laffer (1970), and others.

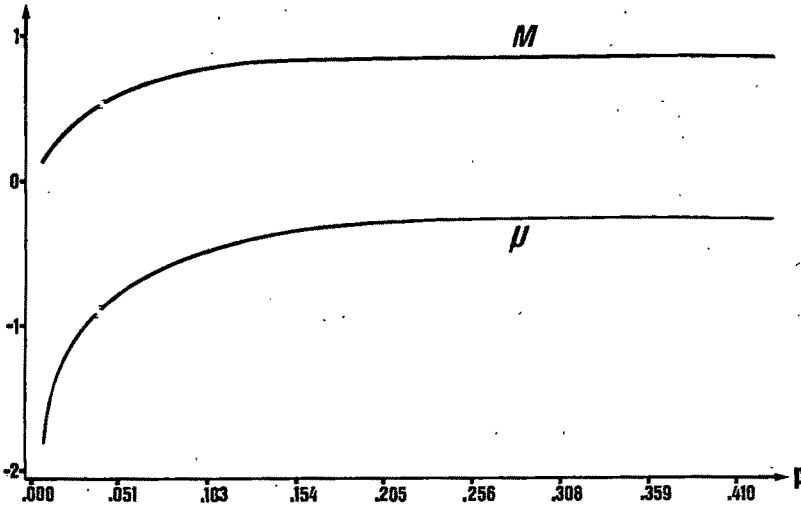


FIGURE 1. PLOT OF TRIGGER AND TARGET POINTS VS. OVERDRAFT RATE

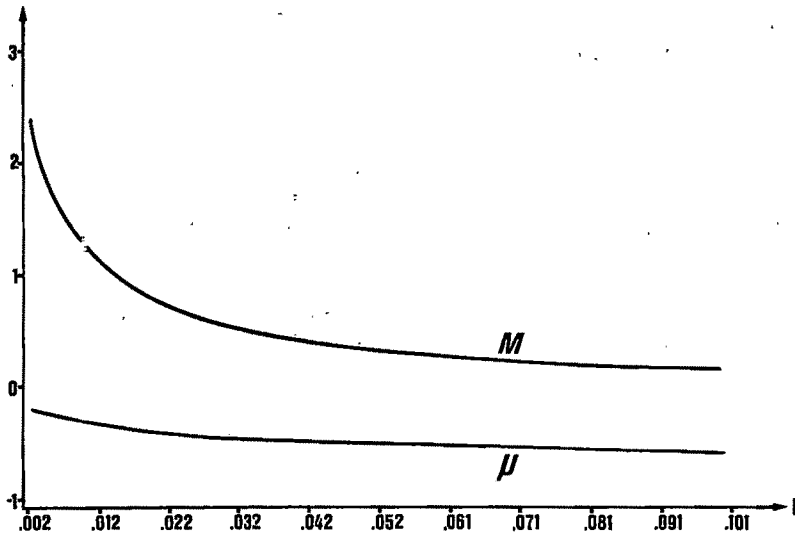


FIGURE 2. PLOT OF TRIGGER AND TARGET POINTS VS. HOLDING RATE

costs in the model are not as crucial as the other two generalizations mentioned previously. When the two kinds of costs are present, the important one is the fixed cost; then the optimal money holding rule is the trigger-target rule that is typical of models with lumpy transactions cost.

B. Very High Overdraft Rate

The most general analysis of money demand when net disbursements include both deterministic and stochastic elements is that of Frenkel and Jovanovic (1980). Their model is a special case of mine with

no proportional cost and with one interest rate instead of three; that is, $c = 0$, $p \rightarrow \infty$, and $r = \alpha$. Assuming $c = 0$ and $p \rightarrow \infty$, I can concentrate on the importance of the generalization $r \neq \alpha$. In this case, equations (16) and (17) yield the following approximate solutions for M and μ :¹⁵

$$(20) \quad M \approx \mu - \frac{K\alpha}{r} + \frac{p}{r} \lambda_2 \mu^2$$

$$(21) \quad e^{-\lambda_1 M} \approx 1 - \lambda_1 \mu - \frac{p}{r} \lambda_1 \lambda_2 \mu^2.$$

Equations (20) and (21) give the following solution for M :

$$(22) \quad e^{-\lambda_1 M} \approx 1 - \lambda_1 M - \frac{K\alpha\lambda_1}{r}.$$

Following Frenkel and Jovanovic, I expand equation (22) in Maclaurin series and ignore terms of third and higher order. The result is¹⁶

$$(23) \quad M \approx \left(\frac{-2K\alpha}{\lambda_1 r} \right)^{1/2}$$

Substituting for λ_1 , then,

$$(24) \quad M \approx \left(\frac{2K\alpha\sigma^2}{r[(g^2 + 2\alpha\sigma^2)^{1/2} - g]} \right)^{1/2}$$

and the approximate solution for μ is

$$(25) \quad \mu \approx - \left(\frac{\lambda_2 \left(\frac{K\alpha}{r} + M \right)}{\frac{p}{r}} \right)^{1/2} \rightarrow 0.$$

¹⁵Equations (20) and (21) are derived by expanding equations (16) and (17), respectively, in Maclaurin series and ignoring terms of third and higher order for those expressions that multiply p/r and of second and higher order for those expressions that do not multiply p/r (which, by assumption, is infinitely high).

¹⁶The approximation used in deriving equation (23) might be less accurate for finite p than that used in equations (20) and (21) because μ , but not M , approaches zero when $p/r \rightarrow \infty$.

Assuming that $r = \alpha$, equation (24) reduces to the central result, the money demand equation, of Frenkel and Jovanovic.¹⁷ Equation (24) can serve as a very convenient instrument in examining the significance of constraining the two interest rates to be equal. In fact, one can define at least four different elasticities which are of interest.

(a) Interest elasticity with $r = \alpha$. In this case, I assume that no interest is paid on money; any change in the market interest rate results in an identical change in the cost of holding money. The empirical relevance of this case is to the demand for currency, M_0 . The interest elasticity in this case is the interest elasticity analyzed by Frenkel and Jovanovic, denoted by $\eta(M_0, r)$, and is derived from equation (24) to give

$$(26) \quad \eta(M_0, r) = \left(-\frac{1}{2} \right) (r\sigma^2) \left[\mu^2 + 2r\sigma^2 - \mu(\mu^2 + 2r\sigma^2)^{1/2} \right]^{-1}$$

which satisfies $-1/2 \leq \eta(M_0, r) < 0$.

(b) Elasticity with respect to the market interest rate, α , given a constant holding rate, r . In this case, the interest paid on interest-bearing money, such as NOW accounts, varies with the market interest rate point by point. For instance, when $\alpha = 7$ percent, the NOW rate is 5 percent, and when α increases to 7.5 percent, the NOW rate increases to 5.5 percent so that $r = 2$ percent without change. In this case

$$(27) \quad \eta^{(1)}(M, \alpha) = \frac{1}{2} + \eta(M_0, r)$$

which is nonnegative. Hence, when the cost of holding money is constant, an increase in the market interest rate leads to an increase in money demand. The intuition behind this

¹⁷Frenkel and Jovanovic call equation (24) with $r = \alpha$ "the optimal money holding." Although M is, in general, different from average money holding, the justification for this statement in their model is that M is the only control variable.

apparently surprising result is simple: an increase in the nominal interest rate with no change in the nominal rate on money holding is in fact a decrease in the effective cost of holding money.¹⁸

(c) Elasticity with respect to the interest cost of money, r , given a constant market rate, α . In this case, the interest paid on the interest-bearing money changes without a change in the market rate. For example, a fall in the interest paid on NOW accounts, with no other change, results in a higher r with the same α . The interest elasticity is now

$$(28) \quad \eta(M, r) = -\frac{1}{2}$$

which is the well-known prediction of the square-root rule of the Baumol-Tobin model. This is an interesting result. A well-established fact in the theory of money demand is that the crucial assumption in the Baumol-Tobin model is that of deterministic disbursements; that is, $\mu/\sigma \rightarrow \infty$. What I find here is that their result is robust to a stochastic generalization as long as the rate α does not change. The reason for this is the following. What characterizes the Baumol-Tobin analysis is not only the deterministic nature of their model but also the "steady-state" assumption, which in fact means no discounting. Hence, I conclude that the assumption of constant discount rate is pivotal to the Baumol-Tobin model, a conclusion that could not be derived without the distinction between r and α .

(d) Elasticity with respect to the market interest rate, α , given a constant rate on the interest-bearing money, i . This case seems to be empirically plausible for short-run analysis since the interest paid on checking accounts, i , is much less volatile than competitive market rates.¹⁹ Substituting $r = \alpha -$

i , where i is fixed, in equation (24) I get

$$(29) \quad \eta^{(2)}(M, \alpha) = \frac{-i}{2(\alpha - i)} + \eta(M_0, r)$$

which reduces to case (a) when $i = 0$. Notice, however, that the interest elasticity can now be much larger than one-half in absolute value if i is close to α . For example, when $i = 5.25$ percent and $\alpha = 7$ percent, then $\eta^{(2)}(M, \alpha) = -1.5 + \eta(M_0, r)$, which falls in the $(-2, -1.5)$ region. The reason for this high short-run interest elasticity is that a percentage rise in the market rate is translated into a much larger increase in the cost of holding money.

The analysis of interest rate elasticities can be summarized as follows. By retaining the assumptions of no proportional cost ($c = 0$) and a prohibitively expensive overdraft rate ($p \rightarrow \infty$), one can study the effect of relaxing the assumption of zero interest rate paid on money. The implications of this generalization seem to be fairly important. One can now distinguish between different monetary aggregates in their response to interest rate changes. The interest elasticity of currency demand [case (a)], which is the case that is usually investigated in the literature, is negative and larger than $-1/2$. However, the analysis of interest elasticity of interest-bearing assets is much richer and depends on the way in which interest rates vary. In the short run, when the interest paid on checking accounts is fixed, one expects to see a large drop in the demand for these accounts when the market rate increases [case (d)]. However, in the long run, when the interest paid on NOW and similar accounts adjusts to the new, higher interest rates, the drop in the money demand will be milder, and one might even see a rise in demand [case (b)]. Case (c) studies circumstances in which the interest paid on checking accounts is more volatile than market rates. This might have been the case when the legal restrictions on payment of interest on checking accounts were removed in recent years to produce an abrupt drop in the cost of money holding, not fully accompanied by a similar drop in other interest

¹⁸The possibility of positive interest elasticities arises also in Romer's (1986) model.

¹⁹The rigidity of i can arise from institutional rigidities and from the "menu costs" of administering an interest rate change.

rates. In this case, my model predicts a surprising interest rate elasticity of $-1/2$, the Baumol-Tobin result, even in a stochastic framework.

C. Deterministic Disbursements

When net disbursements are deterministic ($\sigma^2 = 0$, $g > 0$), the quasi-variational inequality (13)–(15) leads to a first-order instead of a second-order differential equation for $V(m)$, $m \geq \mu$. The straightforward solution for M and μ , assuming $c = 0$ for simplicity, is given by the following equations:²⁰

$$(30) \quad rM + p\mu = -\alpha K$$

$$(31) \quad re^{\alpha M/g} + pe^{\alpha \mu/g} = r + p.$$

When the overdraft rate p satisfies $p \rightarrow \infty$, the solution for M and μ becomes the Baumol-Tobin result:²¹

$$(32) \quad M = \left(\frac{2gK}{r} \right)^{1/2}$$

$$(33) \quad \mu = - \left(\frac{2gKr}{p^2} \right)^{1/2} - \frac{\alpha K}{p} \rightarrow 0.$$

Assume now that disbursements are deterministic and agents minimize expected cost per unit of time. This assumption of zero discount rate ($\alpha = 0$), the so-called "steady-state approach," results in equations (3) and (4), Section I, as the solution for M and μ .²² Once again, the Baumol-Tobin result is obtained when $p \rightarrow \infty$.

²⁰See Sulem (1986). Notice that when disbursements are deterministic, the solution can be derived by using simple calculus, and it is not necessary to resort to impulse control methods.

²¹Ignoring terms of third and higher order in the Maclaurin series of $e^{\alpha M/g}$ and second and higher order for $e^{\alpha \mu/g}$, since M is much larger than (the absolute value of) μ in this case.

²²Notice that equations (3) and (4) cannot be derived directly from equations (30) and (31) by substituting $\alpha = 0$. This is because the latter case should be solved differently from the $\alpha > 0$ case, as demonstrated by Bather (1966) and Sulem (1986).

D. Numerical Solutions

Comparative static analysis of the general solution for M and μ , given in equations (16) and (17), is straightforward but tedious. Most of the elasticities have, in general, ambiguous signs.²³ It is perhaps more illuminating to present numerical solutions for M and μ as a function of different parameters.²⁴ I concentrate on the effects of changes in the parameters p , r , and α , the most important variables that distinguish this analysis from previous ones. Figure 1 depicts the change of the target level M and the trigger μ versus the overdraft rate p for discount rate $\alpha = 7$ percent, cost of holding money $r = 2$ percent (corresponding to 5 percent interest paid on checking accounts), $c = 0$, $g = \sigma = 1$, and $K = 0.01$. Both M and μ increase with p at similar rates so that the difference $(M - \mu)$ is not very sensitive to p when p is not very small.²⁵ For example, when p increases from 20 percent to 40 percent, $M - \mu$ decreases from 1.141 to 1.094, which gives an arc elasticity of -0.04 . I shall make use of this observation later. Notice also that the convergence toward the limiting values of M [equation (24)] and μ (zero) for $p \rightarrow \infty$, the values that are closely related to the Frenkel-Jovanovic analysis, is fairly slow. When $p = 50$ percent, then $M = 0.865$ and $\mu = -0.219$; even when p rises to the outrageous rate of 200 percent, $M = 0.936$ and $\mu = -0.105$, compared to the values $M = 1.017$ and $\mu = 0$, which correspond to the

²³An exception is the effect of an increase in the fixed cost on the target M . When $c = 0$, then,

$$\frac{dM}{dK} = \left(\frac{\alpha}{r} \right) \frac{e^{-\lambda_1 \mu}}{e^{-\lambda_1 M} - e^{-\lambda_1 \mu}} \geq 0$$

and the corresponding elasticity varies with different parameters and is not fixed at the level $1/2$.

²⁴The applicability of this method is not as narrow as it sounds because of the homogeneity of M and μ as a function of (g, σ, K) . One can thus interpret the numbers for M , μ , g , σ , and K as representing, say, thousands of dollars.

²⁵For very small values of p , the trigger μ increases very rapidly. This is because $\mu \rightarrow -\infty$ when $p \rightarrow 0$, since it is optimal never to sell bonds in this case.

$p \rightarrow \infty$ approximation. Hence, the generalization to finite values of the overdraft rate p does make a difference even if this rate is high.

Figure 2 presents the effect of the rate of money holding r on the trigger and target levels when the parameters are $p = 20$ percent, $\alpha = 7$ percent, $g = \sigma = 1$, $K = 0.1$, and $c = 0$. As expected, both M and μ fall when r rises. However, the elasticity of any of these two variables with respect to the holding rate r is, in general, less than one-half (in absolute value), unlike the case when the overdraft rate $p \rightarrow \infty$. Notice that when $r \rightarrow 0$, then $M \rightarrow \infty$ and $\mu \rightarrow 0$ and the money balance is always positive. On the other hand, when r becomes large relative to p , the target M is negative, resulting in negative money balances held at all times.²⁶

The target and trigger levels are not very sensitive to changes in the market interest rate α . For a wide range of parameters, M and μ are practically constant even when α changes from 1 percent to 17.5 percent. However, it is interesting to note that the unexpected result of a rise in the money demand when the interest rate α increases, found in the approximation for large p [equation (27)], still holds when p is finite: M increases slightly and μ decreases slightly when α rises, such that the difference $M - \mu$ increases, although by a very small amount.

The effects on M and μ of changing the parameters of the Wiener process and the cost function are as follows. An increase in the mean:variance ratio g/σ^2 raises both M and μ , but $M - \mu$ can rise or fall. An increase in the fixed cost K , by contrast, raises the target M and lowers the trigger μ by large amounts. Thus, the difference $M - \mu$ is quite sensitive to changes in the fixed cost. Variations in the proportional cost c have a much milder effect: large increases in c produce slight decreases in M and μ . This numerical analysis thus extends existing comparative-analysis results that are derived by approximate solutions.²⁷

²⁶In Figure 2, I allow $r > \alpha$, which implies negative interest on demand deposits.

²⁷See, for example, Hadley and Whitin (1963) or Blinder (1981).

V. Aggregate Money Demand

One of the results of my model, $\mu \leq 0$, implies that individuals and firms frequently use their available credit. As discussed for the deterministic case in Section I, this suggests at least three different measures of the money stock.

(i) Weighted average of the money stock, when both positive and negative balances are weighted by their relative frequency. This is probably what individuals or firms perceive as their average money holding. If one denotes by $\phi(m, t)$ the probability density function of having a money balance m at time t ,²⁸ then average holding, denoted by $E_1(m)$, will be

$$(34) \quad E_1(m) = \int_{\mu}^{\infty} m \phi(m, t) dm.$$

(ii) Average positive money holding. This definition is the closest to the current definition of money and is given by

$$(35) \quad E_2(m) = \int_0^{\infty} m \phi(m, t) dm.$$

(iii) Average money stock measured by its distance from the trigger level μ . This redefinition of "zero level" of money stock yields

$$(36) \quad E_3(m) = \int_{\mu}^{\infty} (m - \mu) \phi(m, t) dm \\ = E_1(m) - \mu.$$

As long as $\mu \leq 0$, as will always be the case in my framework, the relationship among the three measures will be

$$(37) \quad E_3(m) \geq E_2(m) \geq E_1(m)$$

where the equalities hold for $\mu = 0$ ($p \rightarrow \infty$), the standard case in the literature.

²⁸ $\phi(m, t)$ is a function of the initial money stock. I omit this as an explicit parameter for simplicity of exposition. Similarly, the time subscript is omitted in $E_1(m)$.

The steady-state distribution of the money stock is defined as²⁹

$$(38) \quad \phi(m) = \lim_{t \rightarrow \infty} \phi(m, t).$$

The derivation of $\phi(m)$ for the one-target-one-threshold (μ, M) policy is presented in Appendix 2.³⁰ The result is

$$(39) \quad \phi(m) = \begin{cases} (M - \mu)^{-1} [1 - e^{-\pi(m - \mu)}] & \text{for } \mu \leq m \leq M \\ (M - \mu)^{-1} [e^{-\pi(m - M)} - e^{-\pi(m - \mu)}] & \text{for } M < m \end{cases}$$

where

$$\pi \equiv \frac{2g}{\sigma^2}.$$

Using (39) one gets

$$(40) \quad E_1(m) = \frac{1}{2} \left(M + \mu + \frac{\sigma^2}{g} \right)$$

$$(41) \quad E_2(m) = (M - \mu)^{-1} \times \left[\frac{M^2}{2} + \frac{M}{\pi} + \pi^{-2} (1 - e^{\pi\mu}) \right]$$

$$(42) \quad E_3(m) = \frac{1}{2} \left(M - \mu + \frac{\sigma^2}{g} \right)$$

which reduce to the equivalent measures (5)–(7) for the deterministic case.

²⁹The steady-state distribution is not a function of the initial money stock.

³⁰Actually, Appendix 2 presents the derivation for the more general (μ_1, M, μ_2) rule where μ_2 is a second, upper threshold that triggers a reduction in the money stock to the target M . Only then is the condition $\mu_2 \rightarrow \infty$ applied. Notice also that when $\mu = 0$, equation (39) reduces to equation (24) in the Frenkel and Jovanovic (1980) paper.

It is interesting to compare the standard definition of the money supply, $E_2(m)$, which ignores credit, to the definition $E_3(m)$. The latter, which defines unutilized credit as money, is probably the measure proposed by Keynes (1930) and used by Laffer (1970) for the U.S. economy. There are many aspects to the question of the appropriate monetary aggregate. For example, it is crucial to understand the process generating the monetary base, which is tightly controlled by the Fed, for its effect on the dynamics of the price level. However, if money is defined as the medium of exchange, our analysis suggests that $E_3(m)$ is probably more appropriate.

The liquidity of a particular financial instrument is measured by the pecuniary and nonpecuniary costs of transferring it to cash or demand deposits. In this respect, credit should be included in the same monetary aggregate as currency and checking accounts, that is, M_1 . Having \$500 in a checking account and a credit line of \$1,000 yields the same amount of means of payment as \$1,500 in one's account, with no credit. The transition from a positive to a negative balance in an account with overdrafting provision is completely smooth and does not involve any cost. My analysis gives this prediction a somewhat more solid theoretical basis. Demand deposits and credit are very close substitutes; depending on the relative costs, optimizing agents replace one with the other with no significant change in their purchasing power. For example, easier availability of credit, represented by lower p in our model, induces almost perfect substitution of credit for demand deposits (M and μ decrease by very similar amounts). This is also the conclusion if one uses $E_3(m)$ as the aggregate money supply, but not $E_2(m)$. This effect seems to be analogous to the rise of bank demand deposits leading to a reduction in the demand for currency.

This is a potentially useful insight for the puzzle of the missing money (Goldfeld, 1976). The official money stock, $E_2(m)$, which treats the trigger level μ as fixed at zero level, is erroneously perceived to fall with extension of credit. The difference between $E_3(m)$ and $E_2(m)$ tends to be quite

significant. For instance, when $\alpha = 7$ percent, $r = 2$ percent, $p = 20$ percent, $\mu = \sigma = 1$, $K = 0.01$ and $c = 0$, one gets $M = 0.780$ and $\mu = -0.362$ to give $E_2(m) = 0.721$, but $E_3(m) = 1.071$. A similar percentage of discrepancy arises for a wide range of parameters.

The model also suggests a similar explanation to the seemingly excess variability of the velocity of money. Availability of credit will lower both the trigger μ and the target M by similar amounts without changing the velocity of money. If the quantity of money is measured by $E_3(m)$, this will be the conclusion. However, when the standard definition $E_2(m)$ is used, the velocity seems to rise.

To conclude, my model suggests that proper definition of the medium of exchange has to include some measure of approved credit lines available to the public. The exact nature of this measure depends on the specific way in which credit is extended. In some countries, for example Britain and Israel, banks set a limit on the overdraft facilities of firms and individuals; and these should be included in the money supply data. Apparently, the overdraft rate within the limit is relatively low, while the penalty for overdrafting above the limit is very high; in this case, the trigger level μ chosen by customers is probably identical to the limit set by the bank, and the inclusion of unutilized overdraft facilities in the money stock corresponds to the $E_3(m)$ definition. Similarly, the terms of trade credit are often set such that μ coincides with the credit limit. Alternatively, additional data, probably by sampling the credit history of firms and individuals, can help determine the chosen credit limits.³¹ Some indirect measure of $E_3(m)$ can be obtained from data on the standard money supply $E_2(m)$. For example, for the deterministic case we get [from equations (4), (6), and (7)]

$$(43) \quad E_3(m) = \left(\frac{r+p}{p} \right)^2 E_2(m).$$

VI. Concluding Remarks

The current work can be extended in several ways. One way would be to allow two overdraft rates, a low one within a limit and a higher rate for overshooting. Another possibility is to allow for a second, higher trigger point that induces a reduction of the money stock when it is "too high," as in Miller and Orr (1966). A different angle of this problem is the study of monetary policy with credit, whether utilized or not. My model suggests that either credit or monetary control can affect the price level, but when applied separately, their effectiveness is limited by the substitutability of credit and demand deposits.

APPENDIX 1

This appendix presents the solution to the quasi-variational inequality (13)–(15). Since the trigger-target (S, s) rule is the optimal policy for this problem, $V(m)$ can be defined over two regimes. When the money stock is below the trigger point μ , it is increased to the target level M , and equation (14) holds as an equality:

$$(A1) \quad V(m) = K + c(M - m) + V(M) \quad \text{for } m < \mu.$$

When $m \geq \mu$, no financial transaction is made, and equation (13) holds as an equality. In this case, the solution of the linear differential equation (13) is

$$(A2) \quad V(m) = -\frac{r}{\alpha} \left(\frac{g}{\alpha} - m \right) + D_1 e^{\lambda_1 m} + D_2 e^{\lambda_2 m} \quad \text{for } m \geq 0$$

$$(A3) \quad V(m) = \frac{p}{\alpha} \left(\frac{g}{\alpha} - m \right) + E_1 e^{\lambda_1 m} + E_2 e^{\lambda_2 m} \quad \text{for } m < 0.$$

The first term in (A2) and (A3) is a particular solution of the nonhomogeneous part of equation (13), and the last two terms are the general solution to the homogeneous part. D_i and E_i ($i = 1, 2$) are constants to be determined, and the roots of the characteristic equation, λ_1 and λ_2 , are given by

$$(A4) \quad \lambda_1 = \sigma^{-2} \left[- (g^2 + 2\alpha\sigma^2)^{1/2} + g \right] \leq 0$$

$$(A5) \quad \lambda_2 = \sigma^{-2} \left[(g^2 + 2\alpha\sigma^2)^{1/2} + g \right] \geq 0.$$

Assume initially that $\mu \leq 0$. In this case, equation (A3) describes the expected cost $V(m)$ for $\mu \leq m < 0$.

³¹The Bank of Israel uses a sample of 400 companies to study the liquidity position.

Complete characterization of the solution requires finding the values of D_1 , D_2 , E_1 , E_2 , μ , and M . These six parameters are solved using the following six conditions.

(a) Continuity at $m=0$:

$$(A6) \quad -\frac{rg}{\alpha^2} + D_1 + D_2 = \frac{pg}{\alpha^2} + E_1 + E_2.$$

(b) Continuous derivative at $m=0$:

$$(A7) \quad \frac{r}{\alpha} + \lambda_1 D_1 + \lambda_2 D_2 = -\frac{p}{\alpha} + \lambda_1 E_1 + \lambda_2 E_2.$$

(c) Continuity at $m=\mu$:

$$(A8) \quad \frac{p}{\alpha} \left(\frac{g}{\alpha} - \mu \right) + E_1 e^{\lambda_1 \mu} + E_2 e^{\lambda_2 \mu} = K + c(M - \mu) + V(M).$$

(d) Continuous derivative at $m=\mu$:

$$(A9) \quad -\frac{p}{\alpha} + \lambda_1 E_1 e^{\lambda_1 \mu} + \lambda_2 E_2 e^{\lambda_2 \mu} = -c.$$

(e) M is the optimal target. Optimizing over M in equation (A1) yields

$$(A10) \quad V'(M) = -c.$$

(f) $V(m)$ grows linearly at a rate r/α when $m \rightarrow \infty$:

$$(A11) \quad \lim_{m \rightarrow \infty} V'(m) = \frac{r}{\alpha}$$

which gives immediately

$$(A12) \quad D_2 = 0.$$

The rest of the parameters are

$$(A13) \quad D_1 = \frac{1}{\lambda_1} e^{-\lambda_1 \mu} \left(\frac{p}{\alpha} - c \right) + \frac{1}{(\lambda_1 - \lambda_2)} \left(\frac{p+r}{\alpha} \right) \left[\frac{\lambda_2}{\lambda_1} e^{(\lambda_2 - \lambda_1)\mu} \right]$$

$$(A14) \quad E_1 = \frac{1}{\lambda_1} e^{-\lambda_1 \mu} \left(\frac{p}{\alpha} - c \right) - \frac{1}{(\lambda_1 - \lambda_2)} \left(\frac{p+r}{\alpha} \right) e^{(\lambda_2 - \lambda_1)\mu}$$

$$(A15) \quad E_2 = \frac{\lambda_1}{\lambda_2(\lambda_1 - \lambda_2)} \left(\frac{p+r}{\alpha} \right)$$

The solution for the target M and the trigger μ is given by the following equations:

$$(A16) \quad M = (\alpha c + r)^{-1} \times \left[(\alpha c - p)\mu + \left(\frac{p+r}{\lambda_2} \right) (e^{\lambda_2 \mu} - 1) - \alpha K \right]$$

$$(A17) \quad e^{-\lambda_1 M} = (\alpha c + r)^{-1} \left[(\alpha c - p)e^{-\lambda_1 \mu} + \left(\frac{p+r}{\lambda_1 - \lambda_2} \right) (\lambda_1 e^{(\lambda_2 - \lambda_1)\mu} - \lambda_2) \right]$$

It is straightforward to see that the analogue of equations (A6)–(A11) for the case $\mu > 0$ yields the solution $\mu = M$, which would lead to infinite expected costs and which could, therefore, not be optimal. I thus conclude that equations (A16) and (A17) determine the optimal levels of the target money level M and the trigger point μ and that μ satisfies $\mu \leq 0$.

APPENDIX 2

In this appendix I derive equation (39) and present the steady-state distribution of a Brownian motion with a drift when there are two thresholds, μ_1 and μ_2 , and one target, M . The derivation is based on Chapter 15 of Karlin and Taylor (1981). They show (section 8, E) that (μ_1, M, μ_2) diffusion process with mean g and variance σ^2 converges to the following limiting distribution:

$$(A18) \quad \phi(m) = \lim_{t \rightarrow \infty} \phi(m, t) = G(M, m) \int_{\mu_1}^{\mu_2} G(M, y) dy$$

where $G(x, y)$ is the Green function of the diffusion process defined by

$$(A19) \quad G(x, y) = \begin{cases} \frac{2[S(x) - s(\mu_1)][S(\mu_2) - S(y)]}{\sigma^2 s(x)[S(\mu_2) - S(\mu_1)]} & \text{for } \mu_1 \leq x \leq y \leq \mu_2 \\ \frac{2[S(\mu_2) - S(x)][S(y) - S(\mu_1)]}{\sigma^2 s(x)[S(\mu_2) - S(\mu_1)]} & \text{for } \mu_1 \leq y \leq x \leq \mu_2 \end{cases}$$

and where $S(\cdot)$ and $s(\cdot)$, for a Brownian motion with a drift, can be expressed as follows (Karlin and Taylor, p. 205):

$$(A20) \quad s(x) = \exp(-2gx/\sigma^2)$$

$$S(x) = Ax + B \quad (A \text{ and } B \text{ are constants}).$$

The integration required in the denominator of equation (A18) can be performed straightforwardly to get

$$\begin{aligned}
 (A21) \quad & \int_{\mu_1}^{\mu_2} G(M, y) dy \\
 &= \int_{\mu_1}^M G(M, y) dy + \int_M^{\mu_2} G(M, y) dy \\
 &= 2AN / \{ [s(\mu_2) - s(\mu_1)] \sigma^2 s(M) \}
 \end{aligned}$$

where N is defined by

$$\begin{aligned}
 (A22) \quad N &= (\mu_1 - \mu_2) s(\mu_1) s(\mu_2) \\
 &+ s(M) [s(\mu_1)(M - \mu_1) + s(\mu_2)(\mu_2 - x)].
 \end{aligned}$$

The division of equation (A19) by (A21) yields the following solution for the steady-state distribution:

$$\begin{aligned}
 (A23) \quad \phi(m) &= \begin{cases} (1/N)[s(M) - s(\mu_2)][s(\mu_1) - s(m)] & \text{for } \mu_1 \leq m \leq M \\ (1/N)[s(\mu_1) - s(M)][s(m) - s(\mu_2)] & \text{for } M \leq m \leq \mu_2. \end{cases}
 \end{aligned}$$

Since I am interested in the case of no upper boundary, $\mu_2 \rightarrow \infty$, I get from equation (A22)

$$(A24) \quad \lim_{\mu_2 \rightarrow \infty} N = s(\mu_1) s(M) (M - \mu_1)$$

and from equation (A23)

$$\begin{aligned}
 (A25) \quad \lim_{\mu_2 \rightarrow \infty} \phi(m) &= \begin{cases} \frac{s(\mu_1) - s(m)}{s(\mu_1)(M - \mu_1)} & \text{for } \mu_1 \leq m \leq M \\ \frac{s(m)[s(\mu_1) - s(M)]}{s(\mu_1) s(M) (M - \mu_1)} & \text{for } M \leq m \end{cases}
 \end{aligned}$$

which is equation (39) in Section V.

REFERENCES

- Bather, J. A., "A Continuous Time Inventory Model," *Journal of Applied Probability*, December 1966, 3, 538-49.
- Baumol, William, J., "The Transactions Demand for Cash—An Inventory Theoretic Approach," *Quarterly Journal of Economics*, November 1952, 66, 545-56.
- Bensoussan, A. and Lions, J. L., *Contrôle Impulsionnel et Inéquations Quasi-Variationnelles*, Paris: Dunod, 1982.
- Blinder, Alan, S., "Retail Inventory Behavior and Business Fluctuations," *Brookings Papers on Economic Activity*, No. 2, 1981, 443-520.
- Constantinides, George M. and Richard, Scott F., "Existence of Optimal Simple Policies for Discounted-Cost Inventory and Cash Management in Continuous Time," *Operations Research*, July-August 1978, 26, 620-36.
- Dixit, Avinash, "A Simplified Exposition of the Theory of Optimal Control of Brownian Motion," mimeo, Princeton University, June 1989.
- Eppen, Gary D. and Fama, Eugene F., "Cash Balance and Simple Portfolio Problems with Proportional Costs," *International Economic Review*, June 1969, 10, 119-33.
- Frenkel, Jacob A. and Jovanovic, Boyan, "On Transactions and Precautionary Demand for Money," *Quarterly Journal of Economics*, August 1980, 94, 24-43.
- Goldfeld, Stephen M., "The Case of the Missing Money," *Brookings Papers on Economic Activity*, No. 3, 1976, 683-730.
- Hadley, G. and Whitin, T. M., *Analysis of Inventory Systems*, Englewood Cliffs, NJ: Prentice-Hall, 1963.
- Kanninen, V., "The Role of Bank Overdrafts in the Finnish Financial System," *Empirical Economics*, 1978, 3(1), 31-47.
- Karlin, Samuel and Taylor, Howard M., *A Second Course in Stochastic Processes*, New York: Academic Press, 1981.
- Keynes, John Maynard, *A Treatise on Money*, Vol. I, London: Macmillan, 1930.
- Laffer, Arthur B., "Trade Credit and the Money Market," *Journal of Political Economy*, April 1970, 78, 239-67.
- Lavington, Frederick, *The English Capital Market*, London: Frank Cass, 1921, reprinted New York, 1968.
- Lucas, Robert E. and Stokey, Nancy L., "Optimal Fiscal and Monetary Policy in an Economy Without Capital," *Journal of Monetary Economics*, July 1983, 12, 55-93.
- Milbourne, R. D., Buckholtz, P. and Wasan, M. T., "A Theoretical Derivation of the Functional Form of Short Run Money Holdings," *Review of Economic Studies*, July 1983, 50, 531-41.

- Miller, Merton H. and Orr, Daniel, "A Model of the Demand for Money by Firms," *Quarterly Journal of Economics*, August 1966, 80, 413-35.
- Romer, David, "A Simple General Equilibrium Version of the Baumol-Tobin Model," *Quarterly Journal of Economics*, November 1986, 101, 663-85.
- _____, "The Monetary Transmission Mechanism in a General Equilibrium Version of the Baumol-Tobin Model," mimeo, Princeton University, 1987.
- Sulem, Agnes, "A Solvable One-Dimensional Model of a Diffusion Inventory System," *Mathematics of Operations Research*, February 1986, 11, 125-33.
- Tobin, James, "The Interest Elasticity of the Transactions Demand for Cash," *Review of Economics and Statistics*, August 1956, 38, 241-47.

Tax Smoothing with Financial Instruments

By HENNING BOHN*

The paper analyzes the optimal structure of government debt in a stochastic environment. In a model with distortionary taxes, the government should smooth tax rates over states of nature as well as over time. Government liabilities should be structured to hedge against macroeconomic shocks that affect the government budget. The optimal structure of government liabilities generally includes some "risky" securities which are state-contingent in real terms. The empirical part of the paper tests for tax smoothing and then studies state contingencies implemented by some specific securities including nominal debt, long-term bonds, equity, and foreign-currency debt. (JEL 321)

The United States government issues Treasury bonds and bills of various maturities. They are considered risk-free in terms of default risk, though their real value may fluctuate considerably. This paper is concerned with questions of what may motivate such a debt structure and whether it is an optimal one.

The paper analyzes government policies that maximize welfare. The welfare-maximizing approach of analyzing government debt policy was introduced by Robert Barro (1979). He shows that, in a deterministic environment, optimal tax and debt policy should smooth tax rates over time. Optimal policy in a stochastic environment calls for state-contingent tax rates that must be supported by state-contingent debt, as Robert Lucas and Nancy Stokey (1983) have shown. I consider a stochastic version of Barro's model. Key assumptions (discussed in more detail below) are that welfare losses due to distortionary taxation can be summarized by a convex function of tax rates and that asset prices and return distributions are exogenous.

Optimal policy will then smooth tax rates over time and over states of nature. If there are macroeconomic shocks that affect the

government budget, government liabilities should provide a hedge against these shocks. This characterizes the optimal structure of government liabilities.

If markets are complete and if the government can trade on all markets, the model has the very strong implication that changes in tax rates should never occur, because government liabilities could be contingent on all conceivable shocks. However, markets may be incomplete for various informational or incentive reasons, and even if they were complete, the government may not be able to operate on all markets.¹ Definitive answers on why certain markets are missing

¹See Franklin Allen and Douglas Gale (1988) on market incompleteness and Gale (1990) on the possibility that innovative debt management may open new markets. Here, the critical issue is not missing markets per se, but government access to markets. Government activity on markets designed to provide hedges against budgetary uncertainty may be particularly problematic, because of incentive and asymmetric-information problems. For example, the government might be able to reduce the volatility of taxes by issuing securities contingent on government spending or on the tax rate itself. However, if such securities were issued, the government would have an incentive to manipulate their payoffs by changing spending or taxes. One cannot claim, though, that incentive problems provide a full explanation for imperfect tax smoothing, since nominal debt is traded even though it clearly creates a time-consistency problem. Gale's point that innovative debt management may improve welfare is very much consistent with this paper (see Section IV-C), but to be cautious, I will only consider currently existing securities.

*Department of Finance, the Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6367, I thank all the members of the macro-lunch group at the University of Pennsylvania and three anonymous referees for many valuable comments.

and on whether or not the government may trade on a certain market are beyond the scope of this paper. Nonetheless, unless one can rule out any incompleteness or incentive problems, it seems too much to require that the government stabilize tax rates perfectly. Therefore, I take the set of available securities (bundles of state-contingent claims) considered for the government portfolio as exogenous.

For any given set of securities, one can compute the optimal portfolio of government liabilities formed with these securities. The optimality conditions are that the tax rate is uncorrelated with the return on each security. For several well-known and widely traded securities (short- and long-term dollar bonds, stocks, foreign exchange, and foreign bonds), I test the zero-correlation condition and estimate optimal portfolios, making different assumptions about the set of available securities. The tests provide an assessment of tax smoothing as a positive theory of government. The estimated optimal portfolios in comparison with the actual structure of U.S. government debt show in which direction debt policy should be modified to improve tax smoothing. Since the estimation is more constructive and because of robustness considerations, the emphasis will be on the analysis of optimal portfolios, rather than on testing.

For the United States, several questions about the debt structure are interesting under positive as well as normative aspects. The first question relates to the absence of indexed debt (see, e.g., Stanley Fischer, 1983; Bohn, 1988). Nonexistence may be interpreted as indicating that a risk-free asset cannot be created, or it may be an equilibrium phenomenon, meaning that the government could issue indexed debt if it wanted to. The example of Britain suggests that indexed debt could be issued easily. The paper shows that, indeed, the specific state contingency implemented by relating the real value of government debt to inflation has desirable hedging properties. Thus, nonindexation is consistent with optimal policy.²

A second question is about the maturity structure of debt. In a discrete-time framework, the real value of debt with maturity greater than one period varies with nominal interest rates. I show that this type of contingency is also desirable for a welfare-maximizing government, though the evidence is weaker than that in favor of nominal debt.

Third, one may ask whether the government can improve on the current practice of issuing only domestic debt securities. Though an exhaustive survey of alternatives is beyond the scope of this paper, the two cases of stocks and foreign-currency debt are explored. One finds that the government could indeed hedge against shocks due to cyclical fluctuations by taking a short position in the stock market. Movements in some foreign interest rates also have desirable hedging properties, though there is little support for taking exchange-rate risk. Overall, it seems that the government could improve welfare by looking outside the set of dollar-dominated debt securities in structuring its liabilities.

The results on debt management rely heavily on the optimality of tax smoothing. Though the notion that excess burden increases on the margin with increasing tax rates is familiar from atemporal models of taxation, tax smoothing is not always optimal in dynamic models (see Olivier Blanchard and Fischer, 1989 Ch. 11.3; Richard Tresch, 1981 Part III). In general, excess burden may depend on other variables in addition to current tax rates (see, e.g., Lucas and Stokey, 1983). The tax-smoothing approach may work well for economies where taxes are labor income taxes or other taxes with largely static effects, but it is probably less appropriate for cases in which taxation has significant effects on interest rates or on capital accumulation (see Lucas and Stokey [1983] and Kenneth Judd [1989], respectively). In particular, if debt management affected interest rates, the qualitative nature of the government's optimization problem would change significantly, because the

²This abstracts from issues of time consistency, which would favor indexation (Guillermo Calvo, 1978;

Lucas and Stokey, 1983; Bohn, 1988). Time-consistency issues have intentionally been omitted from the paper, because they would distract from the analysis of the government debt portfolio as a whole.

government would no longer behave as a price taker on financial markets. Thus, the normative results of this paper should apply to economies where debt management has few macroeconomic effects and where the current tax rate is the main determinant of excess burden, but they may have to be interpreted more cautiously otherwise.³

A final simplifying assumption is that individuals are risk-neutral. All expected returns are then tied to the rate of time preference, making them exogenous in a straightforward way. Somewhat surprisingly for a paper concerned with hedging, this assumption does not seem to affect the results; but it helps to focus on the government's problem. Regardless of the degree of individual risk aversion, the convexity of excess burden makes the government act "more risk-averse" than taxpayers and implies that tax smoothing is a close approximation to the optimal policy. Since the risk-neutral case shows most clearly how the government's hedging motive derives from the (added) concavity in the welfare function created by distortionary taxes, risk neutrality is imposed throughout (see the end of Section I for more details).

The paper is organized as follows. Section I sets up a simple model that provides a tax-smoothing argument for why governments may want to hedge against economic uncertainty. Tests for tax smoothing are in Section II. Section III derives an equation for the optimal liability structure, and Section IV contains estimates of optimal portfolios containing nominal, long-term, and foreign-currency bonds and stocks. Section V summarizes the results.

I. A Framework for Analysis

Barro (1979) has shown that, with distortionary taxes, the government should smooth tax rates over time. This section shows that Barro's approach generalizes in a stochastic environment to tax smoothing over states of nature. The government behaves as if it is

averse to the risk of changing tax rates, even if all individuals are risk-neutral.

I consider a model similar to Barro (1979), except that risky securities are added. In period t , identical, infinitely lived individuals maximize

$$(1) \quad U_t = E_t \sum_{j \geq 0} \rho^j c_{t+j}$$

where $0 < \rho < 1$ is a discount factor and c_{t+j} is consumption in period $t+j$. They own a stream of endowments Y_{t+j} and may trade $K+1$ assets. Let $A_{k,t}$ be the quantity of asset k ($k = 0, 1, \dots, K$) purchased in period t , $p_{t,k}$ be the price of asset k in terms of consumption goods (ex dividend), and $f_{t+j,k}$, $j \geq 1$, be the stream of cash flows (interest payments or dividends) in future periods. Let returns be denoted by $r_{t+1,k} = (p_{t+1,k} + f_{t+1,k})/p_{t,k} - 1$.

Individuals pay taxes on endowments at a rate τ_t . I assume that taxes are distortionary (e.g., because of wasteful efforts of evading or sheltering income). Following Barro (1979), the excess burden of taxation is summarized by a loss function $h(\tau_t)$, which indicates the fraction of endowment "wasted" when taxes are τ_t . Then, the individual budget constraint is

$$(2) \quad c_t + \sum_k p_{t,k} A_{t,k} = Y_t [1 - \tau_t - h(\tau_t)] + \sum_k (p_{t,k} + f_{t,k}) A_{t-1,k}$$

Individual optimization implies asset-pricing equations $p_{t,k} = \rho E_t(p_{t+1,k} + f_{t+1,k})$ for all k . That is, all expected returns must be equal:

$$(3) \quad E_t(1 + r_{t+1,k}) = \frac{1}{\rho} \quad \text{for all } k.$$

It is convenient to introduce several specific securities that will be analyzed later. First, let $k = 0$ be a risk-free (in real terms, i.e., price-level-indexed) one-period security that has a price $p_{t,0} = 1$ and return $r \equiv 1/\rho - 1 > 0$ for all t . Then one can define excess returns $\hat{r}_{t+1,k} \equiv r_{t+1,k} - r$ on assets $k \geq 1$. Individual optimization can be summarized by $E_t \hat{r}_{t+1,k} = 0$.

³At least for the choice of maturities, the assumption that debt management has little macroeconomic effect seems empirically defensible; see Franco Modigliani and Richard Eutch (1966).

Second, I want to discuss assets with returns defined in terms of a nominal unit of account, money. However, I do not want to focus on asset-pricing issues specific to monetary models, nor do I want to discuss optimal monetary policy. While both issues are important topics in themselves, all I need here is a well-defined price level. Therefore, I will just assume that the price level P_t and the rate of inflation, $\pi_t = \log(P_t/P_{t-1})$, follow some stochastic processes. These assumptions could be motivated more rigorously as a limit of a cash-in-advance model with "small" monetary sector.

Third, some securities may be denominated in a foreign currency. Given risk-neutrality of domestic individuals, the closed-economy market-clearing conditions are not essential. That is, the existence of "other" individuals abroad does not change the model significantly. Whenever necessary, I will therefore assume that payoffs of some securities may depend on variables defined within a foreign economy.

The government uses tax revenues $T_t = \tau_t Y_t$ to finance government spending G_t and to service the government debt. I assume that the government can issue arbitrary quantities $D_{t,k}$ of the securities k at the market price. The government budget constraint is

$$(4) \quad T_t = \tau_t Y_t = G_t + \sum_k (p_{t,k} + f_{t,k}) D_{t-1,k} - \sum_k p_{t,k} D_{t,k}.$$

Individual welfare can be written as a function of government policy by substituting (4) and (2) into the individual objective function and dropping irrelevant terms:⁴

$$(5) \quad U_t = E_t \sum_{j \geq 0} \rho^j \{Y_{t+j} [1 - h(\tau_{t+j})]\}.$$

⁴To be exact, $\sum_k (p_{t,k} + f_{t,k})(A_{t-1,k} - D_{t-1,k}) - E_t \sum_{j \geq 0} \rho^j G_{t+j}$ should be added to the right-hand side of (5); but since this is an additive, exogenous term, it does not affect decisions.

The government chooses tax rates and debt structure to maximize (5) subject to (4). In effect, the government objective is to minimize the expected present value of excess burden. The first-order conditions for optimal policy are

$$(6a) \quad E_t[h'(\tau_{t+1})] = h'(\tau_t) \quad \text{for all } k = 0$$

$$(6b) \quad \rho E_t[h'(\tau_{t+1})(1 + r_{t+1,k})] = h'(\tau_t) \quad \text{for all } k > 0.$$

To simplify, assume that excess burden is quadratic; that is, $h(\tau_t) = (h/2)\tau_t^2$ for some $h > 0$. Then (6a) implies tax smoothing over time, $E_t \tau_{t+1} = \tau_t$, as in Barro (1979), or in terms of stationary variables,

$$(7a) \quad E_t(\Delta \tau_{t+1}) = 0.$$

This also determines the path of total debt. Using (3), equation (6b) can be rewritten as

$$(7b) \quad E_t[\Delta \tau_{t+1}(1 + r_{t+1,k})] = 0 \quad \text{for all } k > 0.$$

Notice that this equation—though not (7a)—would hold even if it were impossible to issue a risk-free asset. Combining (7a) and (7b), one obtains

$$(8) \quad E_t(\Delta \tau_{t+1} \hat{r}_{t+1,k}) = E_t(\hat{\tau}_{t+1} \hat{r}_{t+1,k}) = \text{Cov}_t(\tau_{t+1}, r_{t+1,k}) = 0$$

where $\hat{\tau}_{t+1} = \tau_{t+1} - E_t \tau_{t+1}$ is the innovation in tax rates. That is, the government should stabilize taxes across possible states of nature. This implies a zero conditional covariance between taxes and returns on all available securities. The optimality conditions (7) and (8) will be tested in Section II. They implicitly characterize the optimal debt structure, since taxes are a function of debt policy through the budget constraint. This link will be made explicit in Section III.

At this point, a remark on the assumption of risk-neutrality may be appropriate. If preferences (1) were replaced by a time-

additive concave utility function (with money and foreigners excluded for the purpose of this argument), the ratio of $t+1$ and t -dated marginal utilities would enter (3), as in Lucas (1978). Through a similar modification of (5), the same ratio of marginal utilities would enter equations (6)–(8). If one takes such a consumption-based capital-asset-pricing model as the maintained hypothesis for normative analysis, a variation of Rajnish Mehra and Edward Prescott's (1985) argument (based on the fact that consumption growth has low variance) can be used to show that changes in the marginal rate of substitution are quantitatively unimportant (details available from the author). Independently from the Mehra-Prescott argument, the fact that tax rates and consumption have low correlation in U.S. data (less than 0.05 in absolute value for the period 1954–87, using nondurables) suggests that risk-neutrality is a justifiable simplification. Risk aversion does not affect the basic intuition that, because of the convex excess burden in its objective function, the government should smooth tax rates.⁵

II. Optimal Debt

In this section, the first-order conditions for taxes, (7) and (8), are tested for quarterly postwar United States data. The basic sample periods are 1954:2–1987:4 for domestic series and 1973:1–1987:4 for international series. The measure of tax rates is the ratio of federal tax revenue to GNP.⁶

⁵The reference to Mehra and Prescott (1985) leads to a more fundamental problem, however, because their point was that consumption-based asset pricing cannot explain observed risk premiums. If the Mehra-Prescott puzzle means that actual risk premiums are too high for no good reason, one may speculate whether the optimal debt portfolios should be biased toward assets that promise low expected returns. However, if the “true” model that justifies the observed high equity premiums is similar to the Lucas (1978) model in that the same risk factor (perhaps something other than marginal utility of consumption) appears in both consumer and government first-order conditions, the basic intuition will still apply.

⁶The reason for using this measure instead of, say, marginal federal income tax rates is that it is unclear what “the” marginal tax rate is on aggregate. Taxes are

The use of quarterly data seems necessary to obtain reasonably precise estimates of the covariances, but it may be somewhat problematic if tax policy is set less frequently. (The optimal debt portfolios in the next section will turn out to be more robust.) If a quarterly revision of tax policy in light of news about all relevant variables is not feasible in practice (e.g., for technical reasons [tax laws specify annual payments] or because of legislative delays), significant correlations in quarterly data should not be interpreted as rejections of optimality but, rather, as an indication that tax smoothing could be improved, if it were possible to adjust tax laws more frequently. Still, equations (7) and (8) yield the most natural and direct tests of the tax-smoothing model, even if negative conclusions may have to be interpreted cautiously.

The random-walk condition (7a) will only be discussed briefly, since it has been tested before (see Chaipat Sahasakul, 1986; Barro, 1981). Over the postwar sample, the change in tax rates has an insignificant mean of 0.023 percent ($t = 0.59$) and small autocorrelations. The series was regressed on its own lags, on various lagged asset returns, and on monetary data (those used in Section IV). The hypothesis that changes in tax rates are unpredictable was not rejected in any of the tests; the details are therefore omitted. These results confirm Barro (1981), but they are in contrast to the results of Sahasakul, who finds evidence against tax smoothing for a sample period that includes World War II. However, if one looks more carefully, the results do not contradict Sahasakul: simple tax-smoothing models apparently have some problems in explaining World War II data (which is not the subject

levied not just on income but also on other activities; rates differ across individuals; and tax laws contain a multitude of exemptions and exceptions. On aggregate, however, all taxes must be paid out of the available economic resources, GNP. A higher revenue:GNP ratio implies higher tax rates somewhere, no matter how tax laws are structured in detail. Therefore, I consider changes of the revenue:GNP ratio as indicators of changes in tax rates and excess burden. See Barro (1981) for further discussion.

of this paper). On the other hand, (7a) cannot be rejected for the postwar period.⁷

Equation (8) implies that innovations in tax rates should be uncorrelated with all innovations in returns. This is a prediction of the tax-smoothing model that, to my knowledge, has not been tested before. For the implementation of the test, notice that all linear combinations of security returns—differences between pairs of returns in particular—must also be uncorrelated with tax changes and that fixed components of returns do not matter for return innovations. Any real return is the difference of the nominal return and inflation. The innovation in inflation is -1 times the real return innovation of a one-period nominal bond. Thus, provided a one-period nominal bond (a three-month Treasury bill) is in the set of assets, equation (8) implies zero covariance between innovations in tax changes and all nominal returns. Inflation data are not needed. Using linear combinations and dropping known components are also useful for cases in which returns have several parts (e.g., for bonds with different maturities and for foreign-currency investments).

Covariances between innovations of tax rates and returns were computed for a variety of widely available securities. The return series are the change in three-month Treasury bill yields (symbolized DTB), the nominal return on long-term Treasury bond (LRET), nominal stock returns measured by the Standard and Poor's (S&P) 500 index (STOCK), and changes in German and

Japanese exchange rates (EG, EJ), money-market yields (SG, SJ), and long-term government bond yields (LG, LJ).⁸ Yield changes in nominal bonds are used as proxies for capital gains, which are the uncertain part of bond returns.⁹ Exchange rates capture the risky component of the return on a three-month foreign money-market investment. Changes in foreign yields measure the difference between long- and short-term foreign bond-market investments. (Though some of the series represent linear combinations or only the uncertain components of returns, I will refer to them as "returns.") Finally, to test whether indexed debt might improve tax smoothing, the innovation in inflation (GNP-deflator; symbolized P) is considered as a potential return series.

Innovations in returns and tax rates were computed from vector autoregressions (VAR) of tax rates and returns, generally with four lags. For each return series, Table 1 displays the correlation, ρ_k , between innovations in taxes and the return, the covariance denoted by $c_k = E_0(\hat{\tau}_{t+1}\hat{r}_{t+1,k})$ and its asymptotic standard error denoted by $\text{std-}c_k$. In addition, Table 1 indicates (by + symbols) whether an ordinary least-squares regression of tax rates on the current return and lagged values of both series has a significant coefficient on the current return. This test is motivated by the fact that the zero-covariance restriction (8) implies a value of zero for this regression coefficient.

As Table 1 shows, tax smoothing is clearly rejected with both domestic bond series, with stock returns, with the inflation se-

⁷Sahasakul (1986) examines contemporaneous relations between tax rates and other government-sector variables instead of the predictability of tax rate changes, and he uses marginal income-tax rates instead of the revenue:GNP ratio as basic data series. He finds a positive relation between temporary military spending and tax rates, which indicates a violation of equation (7a). Using the data provided in his paper, I find that tax rates are indeed not a random walk over his sample period (1937–82); they are positively autocorrelated. However, for the postwar period (1954–82), (a) the positive relation between temporary military spending and tax rates vanishes, and (b) one cannot reject that tax rates follow a random walk. A Chow test indicates a significant break in the autocorrelation coefficient.

⁸Data sources are national income accounts for macroeconomic series, the International Monetary Fund for international data, the Center for Research in Security Prices (CRSP) for Treasury-bill yields, and Ibbotsen Associates for stock and bond returns. Domestic returns are based on the last day of a quarter, and international returns are based on the last month of a quarter. Details are available from the author.

⁹Because of changing duration, this is only an approximation for finite intervals. For the Treasury-bill market, exact three-month holding returns on six-month bills were available from 1959 on. Estimated results for return innovations computed from yield changes and from holding returns were very similar. Thus, the longer series on yield changes was used throughout the study.

TABLE 1—CORRELATIONS WITH TAX RATES

Return series	ρ_k	c_k	std- c_k
A. 1954:2–1987:4:			
DTB	-0.201	-0.243	0.106****
LRET	-0.281	-0.627	0.206*****
STOCK	-0.196	-0.646	0.289***+
P	-0.294	-0.051	0.016*****
B. 1973:1–1987:4:			
EG	0.091	0.276	0.393
EJ	0.019	0.056	0.377
SG	-0.223	-0.029	0.017*
LG	-0.431	-0.031	0.010*****
SJ	-0.255	-0.027	0.014**+
LJ	-0.141	-0.096	0.088

Legend: ρ_k is the correlation between return series k and the tax rate, c_k is the covariance between return series k and the tax rate, and std- c_k is the asymptotic standard error of c_k .

Notes: Stars indicate rejection of $H_0: c_k = 0$ at the 10-percent (*), 5-percent (**), or 1-percent (***) significance level based on the asymptotic standard error std- c_k . Plus signs (+, ++, or +++) indicate rejection of the same hypothesis at the same, respective, significance levels, based on a regression of tax rates on current returns and four lags of both series.

ries, and with some of the international series. That is, the government would have been able to improve tax smoothing by having different quantities of the corresponding securities in its portfolio of liabilities.¹⁰ Unfortunately, these statistical rejections provide little information on how policy could be improved, and they may be sensitive to institutional rigidities in the tax-setting pro-

cess.¹¹ The next section will therefore explore how the optimal structure of government liabilities can be computed explicitly.

III. The Optimal Structure of Government Liabilities

To obtain a solution for the optimal debt structure, the link between innovations in debt and tax rates must be made explicit (i.e., a formula for innovations in tax rates, $\hat{\tau}_{t+1}$, is needed). Let y_t be the growth in output (= endowments; empirically, GNP), let \bar{y} be the mean, and assume $\bar{y} < r$. Denote the new information about period- $(t + j + 1)$ output growth by $\hat{y}_{t+1+j} = E_{t+1}y_{t+1+j} - E_t y_{t+1+j}$, denote the new information about period- $(t + j + 1)$ government spending relative to output by $\hat{g}_{t+1+j} = (E_{t+1}G_{t+1+j} - E_t G_{t+1+j})/Y_t$, and let $d_{t,k} = p_{t,k}D_{t,k}/Y_t$ be the ratio of security- k debt to output. Then, an approximate solution for the period- $(t + 1)$ innovation in tax rates is

$$\begin{aligned}
 (9) \quad \hat{\tau}_{t-1} &= \tau_{t+1} - E_t \tau_{t+1} \\
 &= (1 - \psi) \cdot \exp(-\bar{y}) \\
 &\quad \times \left[\sum_k \hat{r}_{t+1,k} d_{t,k} + \sum_{j \geq 0} \rho^j \hat{g}_{t+1+j} \right] \\
 &\quad - \tau_t \sum_{j \geq 0} \psi^j \hat{y}_{t+1+j}
 \end{aligned}$$

¹⁰To verify that the results are not due to misspecification of the regressions or the assumed availability of a risk-free asset, a test based on equation (7b) was implemented by testing whether the product series $\Delta\tau_{t+1}(r_{t+1,k} - r_{t+1,l})$ has a zero mean for any pair of security returns (k, l) . [The product $\Delta\tau_{t+1}(1 + r_{t+1,k})$ was not used directly, because its mean is dominated by the $\Delta\tau_{t+1}$ component; a test would differ little from testing (7a).] Series LRET and STOCK, for which complete return data are available, were taken as securities k , the three-month Treasury-bill as l . The t statistics of -3.71 and -2.00 confirm the rejections reported in Table 1 at the same levels of significance (1 percent and 5 percent, respectively). Since the simple tests already yield strong rejections, more elaborate tests (e.g., following Lars Hansen and Kenneth Singleton [1983]) seem unnecessary.

¹¹It may be tempting to use the correlations in Table 1 to assess the economic significance of the rejections. The fact that most correlations are below 0.3 in absolute value (except for LG) suggests that the optimal supply of any one security would have reduced the variance of tax rates by less than 10 percent (18.6 percent for LG; 11.7 percent for all four securities in Part A of Table 1 jointly, estimated using a four-lag VAR with tax rates and all four return series). However, as noted earlier, one should be cautious in interpreting the results based on high-frequency tax-rate data. If there are short-term institutional rigidities in tax policy, marginal changes in the liability portfolio (to optimize debt management) might not translate into changed tax policy on a quarterly basis. The tests show that the variance of tax rates has not been minimized, but they do not reveal what is responsible for the excessive variance.

where $0 < \psi = \rho \cdot \exp(-\bar{y}) < 1$ is a discount factor.

The derivation is conceptually straightforward but lengthy.¹² The idea is that the present value of tax revenues must cover initial debt plus the present value of spending. Any innovation in current or future government spending and any unexpected change in the value of debt therefore forces the government to adjust tax revenues eventually. Because of tax smoothing, a fraction $(1 - \psi)$ of the adjustment takes place immediately. In addition, since the present value of tax revenues depends on the path of output, tax rates have to be changed whenever new information about current or future output is received. As a result, tax rates are increased if the value of debt increases unexpectedly, if estimates of future government spending are revised upwards, or if output is lower than expected.

It may be worth noting that this argument applies even if tax rates cannot be adjusted every period. Then, optimal debt policy would have to stabilize the Lagrange multiplier of the budget constraint, which would replace the marginal excess burden $h'(\tau)$ in the first-order condition (8). This multiplier would be perfectly correlated with the right-hand side of (9), leaving the results for optimal debt structure unchanged.

Using the tax-rate formula (9) in the first-order condition (8), one obtains a system of K equations for the K "risky" securities, $d_{t,k}$, issued by the government:

$$\begin{aligned} & \sum_l \text{Cov}_t(\hat{r}_{t+1,l}, \hat{r}_{t+1,k}) d_{t,l} \\ & + \text{Cov}_t\left(\hat{r}_{t+1,k}, \sum_{j \geq 0} \rho^j \hat{g}_{t+1+j}\right) \\ & - w_t \text{Cov}_t\left(\hat{r}_{t+1,k}, \sum_{j \geq 0} \psi^j \hat{y}_{t+1+j}\right) = 0 \end{aligned}$$

for all l , where $w_t = [\exp(\bar{y})/(1 - \psi)]\tau_t$ is a

weighting factor. To simplify, let \mathbf{d}_t be the vector of risky government debt securities $d_{t,l}$, let Σ_r be the variance-covariance matrix of returns, assumed to be nonsingular, and let $\Sigma_{g,r}$ and $\Sigma_{y,r}$ be the vectors of covariances between returns and the present-value expressions $\sum_{j \geq 0} \psi^j \hat{y}_{t+1+j}$ and $\sum_{j \geq 0} \rho^j \hat{g}_{t+1+j}$, respectively. Then, the above equation can be restated as $\Sigma_r \mathbf{d}_t + \Sigma_{g,r} - w_t \Sigma_{y,r} = 0$, and the optimal debt structure is

$$(10) \quad \mathbf{d}_t = w_t \Sigma_r^{-1} \cdot \Sigma_{y,r} - \Sigma_r^{-1} \cdot \Sigma_{g,r}$$

Thus, the general formula for optimal debt structure involves covariances of returns with innovations in output and government spending.

Output (or equivalently, aggregate income) matters in this model, because it forms the tax base and because high tax rates cause distortions. Given desired levels of revenues, tax rates must increase if output falls. Since tax rates are smoothed over time, any news about future output is also relevant. Consequently, permanent changes in output have much larger effects than temporary changes. Unexpectedly high government spending has an additional effect on tax rates. The optimal debt structure is chosen to hedge against these sources of uncertainty and thereby minimizes fluctuations in tax rates.

IV. Estimates of the Optimal Debt Structure

In this section, formula (10) will be used to explore the optimal structure of government liabilities.

A. Methodology

Equation (10) identifies uncertain output and uncertain government spending as sources of risk. Since the two covariance vectors enter as weighted sums in this equation, the optimal debt structure can be interpreted as the sum of two components. A priori, it seems likely that output variation is a quantitatively significant source of risk; the cyclical volatility of budget deficits is

¹²Details are available from the author. Quadratic excess burden $h(\tau)$ is assumed, and the growth rate of output, y_t , should be stationary but not necessarily independently and identically distributed.

well documented. Interesting correlations of returns with output (GNP) were indeed found, while correlations with spending turned out to be small and largely insignificant. (Details are in an earlier version of this paper, which is available on request.) In reporting empirical results, I will therefore focus on the output component. This is done formally by imposing the auxiliary assumption that correlations between government spending and debt are approximately zero, $\Sigma_{g,r} = 0$. Then, the optimal structure of debt is proportional to the vector $s \equiv \Sigma_r^{-1} \cdot \Sigma_{y,r}$, which depends only on the variance-covariance matrix of innovations in output and security returns, $d_t = w_t s$.

It is useful to keep the factor of proportionality, w_t , separate because w_t depends critically on the discount rate applied to future output. For a real discount rate of $1 - \psi = 1$ percent per quarter, $\tau_t \equiv 0.2$, and $\exp(\bar{y}) \equiv 1$, the proportionality factor is $w_t = [\exp(\bar{y}) / (1 - \psi)] \tau_t \equiv 20$; but if $1 - \psi = 0.5$ percent were assumed, the factor would double. Recalling that $\Sigma_{y,r}$ is the covariance between the *present value* of output and returns, the discount factor also enters into the vector s , but it turns out that different discount rates have a negligible effect on the estimates. (To save space, results for s will be shown for $\psi = 0.99$ only.)

Thus, I will concentrate on computing the vector s , which indicates whether securities enter with positive or negative sign into the optimal portfolio and in which relative quantities. For the interpretation, a range of "reasonable" weights, say between 10 and 40, may be applied to compute d_t . The main econometric problem in estimating $\Sigma_{y,r}$ or s is to identify the innovations, in particular the change in expectations of "far out" realizations of output growth, which enter $\Sigma_{y,r}$ through the expected-present-value expression $\sum_{j \geq 0} \psi^j \hat{y}_{t+1+j}$. Vector-autoregression (VAR) techniques were used because they seem ideally suited for the tasks of extracting the covariance structure and computing projections of a multivariate process.

For all securities, I started with a minimal bivariate VAR including only GNP growth (for y_t) and a single return series, using

quarterly U.S. data from 1954:2 to 1987:4, including a constant and four lags. Several alternative versions were estimated to see whether the results are robust to changes in specification.¹³ Finally, I computed optimal debt structures for several return series simultaneously, with and without additional information variables. These alternative specifications will only be displayed when they affect the conclusions from the basic bivariate process.

Consistent point estimates and asymptotic standard errors of the elements of $\Sigma_{y,r}$ and s (denoted by c_k and s_k in the tables) can be obtained as functions of the VAR coefficients and the residual variance-covariance matrix (see Theodore Anderson, 1958; Takeshi Amemiya, 1985; Peter Schmidt, 1976). If the VAR includes only y_t and a single return series, the signs of c_1 and h_1 can also be determined with a simpler auxiliary regression of y_t on the current return and lagged values of both series. Details are available from the author upon request.

B. Domestic Debt Securities

Currently all U.S. government liabilities are nonindexed dollar-denominated debt securities with various maturities. To focus on indexation first, consider a debt portfolio with only two securities, an indexed bond ($k = 0$) and a one-period nominal bond ($k = 1$). Since the real return on a one-period nominal bond is the known promised yield, the innovation in real return is -1 times the rate of inflation. Thus, the covariance between the present value of GNP and inflation determines whether nominal debt is desirable as a hedge. Estimates based on VAR's with GNP growth and inflation are displayed in Table 2 (where all estimates

¹³The alternative estimates used eight instead of four lags, sample periods 1954:2–1972:4 and 1973:1–1978:4 (intended to capture potential breaks in the processes), or one or more additional variables in the information set (other returns, military spending, money supply M1, the monetary base, and import prices). All macroeconomic variables are log-differenced.

TABLE 2—RETURNS ON NOMINAL BONDS

VAR	ρ_k	c_k	std- c_k	s_k	std- s_k
1	0.853	0.511	0.254***	3.28	1.59**
2	0.534	0.240	0.102**	1.95	0.80**
3	0.540	0.237	0.096**	2.00	0.77***
4	0.461	0.166	0.075**	1.63	0.70**
5	0.388	0.133	0.069*	1.34	0.68**

Legend: ρ_k is the correlation between a return series and the present value of output; c_k is the covariance between a return series and the present value of output; std- c_k is the asymptotic standard error of c_k ; s_k is the indicator of optimal supply of a security, as defined in Section IV-A; and std- s_k is the asymptotic standard error of s_k .

Notes: The columns of c_k and std- c_k have been multiplied by 10^4 to improve readability. Stars indicate rejections of $c_k = 0$ or $s_k = 0$ at the 10-percent (*), 5-percent (**), or 1-percent (***) significance level based on the asymptotic standard errors (Wald tests). Plus signs (+, ++, or +++) indicate rejection of the same hypotheses at the same, respective, significance levels, based on a regression of GNP growth on current returns and four lags of both series. The VAR specifications are:

- 1) bivariate with GNP growth and inflation, sample 1954:2–1987:4, four lags;
- 2) with money supply M1, money base, and DTB as information variables; otherwise as for VAR 1;
- 3) with M1, money base, military spending, and DTB as information variables; otherwise as for VAR 1;
- 4) with M1, money base, import prices, and DTB as information variables; otherwise as for VAR 1;
- 5) with M1, money base, import prices, military spending, and DTB as information variables; otherwise as for VAR 1.

have been multiplied by -1 ; i.e., a positive value means that nominal bonds should be issued).

In the basic specification, VAR 1, the correlation between the innovations is 0.85. This seems extraordinarily high and may reflect the omission of other variables that predict output and inflation; but even if various other variables are included (see VAR's 2–5), the correlations remain around 0.50. The estimates are not only statistically but also economically significant. Taking $s_1 = 1.34$ (the lowest estimate) and a proportionality factor of $w_t = 20$, the ratio of nominal debt (with one-quarter maturity) to GNP should be about $d_1 = 26.8$, as opposed to the current debt:GNP ratio of around 0.50. One has to keep in mind, though, that time-consistency issues are not modeled here, which may reduce the optimal amount of nominal debt. Still, the results suggest that the optimal solution may require the government to hold indexed bonds and to issue nominal debt in an amount far exceeding its total debt. If such simultaneous large long and short positions are impractical, the

current policy of issuing only nominal bonds may be interpreted as a corner solution. Alternatively, it may be that the three-month maturity of nominal bonds implicit in quarterly data is too low.

In analyzing the optimal maturity distribution, I will limit the study to the choice between one-period, two-period, and long-term bonds, represented by three-month Treasury bills, six-month Treasury bills, and the longest-term Treasury bonds. Because of the high correlation between interest rates, more variables would probably not add any new insights.

To obtain implications for observable (meaning nominal) return series and to prevent a repetition of the indexation question, it is convenient to restate the term structure in terms of three-month Treasury bills and forward contracts. Given that nominal debt should be issued (as determined above), the question is only whether some of it should have maturities longer than three months. The real return innovation of the long-term bond relative to the three-month bill is given by its nominal return, LRET. The real re-

TABLE 3—MATURITY CHOICE

VAR	ρ_k	c_k	std- c_k	s_k	std- s_k
A. Changes in Treasury-bill rates:					
1	0.084	0.268	0.590	0.38	0.83
2	0.568	1.553	0.747***++	2.41	1.12**
B. Returns on long-term bonds:					
1	0.155	0.70	1.507	0.039	0.060
2	0.334	1.796	1.397	0.074	0.057

Legend: See Table 2 for notation.

Notes: VAR's 1 are bivariate processes with GNP growth and a return series (DTB or LRET), using sample data for 1954:2–1987:4, including four lags. VAR's 2 use eight lags instead. The columns of c_k and std- c_k have been multiplied by 10^5 in Part A and by 10^4 in Part B to improve readability.

turn on a six-month, relative to a three-month, Treasury bill is the six-month yield at the beginning of the period minus the three-month yield at the end of the period. Since six-month yields are not available for the entire sample, the return innovation is proxied by the change in three-month yields, denoted by DTB (For 1960–87, results were similar to those with the exact return series and are therefore not reported.)

Innovations in LRET and DTB can be interpreted as return innovations on a three-month forward contract on the security; that is, correlations with the present value of output determine whether the government has a hedging demand for forward contracts. An optimal short position together with the supply of three-month Treasury bills would establish the optimality of issuing longer-term bonds.

Estimates are displayed in Table 3. All correlations between the present value of output with DTB and LRET have positive signs, which indicates that two-period or long-term nominal bonds (or equivalently, forward contracts) should be issued. Unfortunately, only the estimate for DTB based on an eight-lag VAR is significant. All the positive correlations seem to be due to a delayed reaction of output to interest rates, which comes out strongest in processes with long lags. Similar results were obtained when the optimal supply of three-month, six-month, and long-term bonds (or two of the three) was estimated jointly as a vector; only the estimate for nominal debt (vs. in-

dexation) was significant. Overall, the point estimates provide some support for issuing long-term debt, but because of the large standard errors, a variety of maturity distributions could be consistent with optimal policy.

C. Nontraditional Government Liabilities

There is no reason why governments should restrict their liabilities to nominal or indexed debt securities. In this section, I will consider two other classes of securities: stocks and foreign-currency debt.

German mark- and Swiss franc-denominated "Carter-bonds" were issued by the U.S. government in 1978. More recently, yen-denominated debt has been proposed. Concentrating on marks and yen, I consider one-period, two-period, and long-term foreign-currency bonds. Their real returns are linear combinations of nominal exchange-rate changes, nominal interest-rate changes, and domestic inflation. As in the domestic context, it is instructive to analyze the components, which may be interpreted as forward contracts.

Results are displayed in Table 4, all for the period of flexible exchange rates 1973:1–1987:4. The return series are the rates of dollar depreciation relative to marks and yen (EG and EJ, respectively) and -1 times the change in short- and long-term German and Japanese interest rates (SG and LG for Germany, SJ and LJ for Japan, respectively).

TABLE 4—OPTIMAL PORTFOLIOS WITH FOREIGN-CURRENCY-DENOMINATED SECURITIES

RET	ρ_k	c_k	std- c_k	s_k	std- s_k
A. One-period U.S. dollar-bonds and one- and two-period mark bonds:					
P	0.796	0.387	0.285	3.264	2.239**
EG	0.324	2.373	2.154	0.079	0.064
SG	0.251	0.076	0.075	2.606	1.587
B. One-period U.S. dollar and mark bonds and long-term mark bonds:					
P	0.361	0.154	0.250	0.626	1.927
EG	0.340	2.286	1.844	0.133	0.960**
LG	0.544	0.081	0.051	7.258	3.215**
C. One-period U.S. dollar-bonds and one- and two-period yen bonds:					
P	0.699	0.291	0.184	2.333	1.515
EJ	-0.191	-1.312	2.094	-0.036	0.063
SJ	0.795	0.188	0.075**	4.657	1.629***
D. One-period U.S. dollar and yen bonds and long-term yen bonds:					
P	0.540	0.226	0.177	1.97	0.364
EJ	0.057	0.355	1.896	0.070	0.076
LJ	0.598	0.091	0.048*	7.528	3.526**

Legend: See Table 2 for notation.

Notes: Each panel is estimated with a four-variable VAR with GNP and the three return series listed under RET, using sample data from 1973:1–1987:4, including four lags. The columns of c_k and std- c_k have been multiplied by 10^4 to improve readability.

Since the question of whether there is an incremental benefit in issuing German or Japanese currency bonds arises in the presence of domestic nominal bonds, I concentrate on the multivariate framework. In Part A of Table 4, nominal domestic bonds (with real returns contingent on $-1 \times$ inflation, P) are considered jointly with one- and two-period German bonds. The quarterly exchange-rate movements, EG , indicate the return on three-month investments in Germany relative to Treasury bills. The change in German interest rates, SG , proxies the return on six-month relative to three-month investments. The positive correlations and point estimates suggest that both variables may have hedging roles, but they are insignificant.

Significant positive results were obtained by the analogous regressions with long-term German bonds in Part B of Table 4 and with short- and long-term Japanese bonds in Parts C and D. (Bivariate VAR's with GNP growth and a single return series were

similar: estimates for LG , SJ , and LJ are significant.) Interestingly, the optimal exposure to yield changes exceeds the optimal total exposure to exchange-rate risk in all cases.¹⁴ The optimal hedge would have to combine short positions in short-term foreign securities with larger holdings of longer-term bonds. Overall, exposure to selected foreign interest rates appears to be desirable. Though a more comprehensive analysis of foreign-currency debt is beyond the scope of this article, there seems to be some potential for improvements in United States debt policy in this direction.

Finally, stock prices are commonly considered to be highly cyclical variables, which makes them natural candidates for hedging output risk. Given that nominal debt is pre-

¹⁴Exposure to exchange-rate risk may be desirable for reasons not considered here (e.g., for providing incentives or credibility in the context of exchange-rate stabilization).

TABLE 5—OPTIMAL PORTFOLIOS OF NOMINAL BONDS AND STOCKS

RET	ρ_k	c_k	std- c_k	s_k	std- s_k
A. Nominal bonds and stocks:					
P	0.847	0.491	0.253**	2.932	1.652**
Stocks	0.451	5.073	2.023	0.050	0.034
B. Nominal bonds and stocks (five-variable VAR):					
P	0.660	0.334	0.150**	2.031	1.035**
Stocks	0.530	5.198	1.876***	0.074	0.033**

Legend: See Table 2 for notation.

Notes: Part A is estimated with a three-variable VAR with GNP, inflation, and stock returns, using sample data from 1954:2–1987:4, including four lags. The VAR for Part B includes M1 and the monetary base as additional variables. The columns of c_k and std- c_k have been multiplied by 10^4 to improve readability.

sent, optimal nominal debt and the optimal stock market position were estimated jointly in Table 5. Part A is based on a VAR with output growth, inflation, and stock returns (measured by the S&P 500 index). Part B is based on a VAR that includes the M1 money supply and the monetary base as additional variables. In both cases, the optimal debt structure includes a short position in stocks, which is significant in the larger process that is based on the better estimate of inflation. (Similar estimates were obtained in a bivariate VAR with GNP growth and stock returns only, which were significant at the 1-percent level.)

To my knowledge, a proposal suggesting government participation in the stock market has not been made before. Based on the statistical evidence linking stocks to the present value of output, a short position would provide a hedge against cyclical shortfalls in government revenue. However, when exploring such nonstandard financing strategies, one may ask why the government should not go one step further and sell synthetic securities that are directly contingent on GNP. A theory of market incompleteness (beyond the scope of this paper; see Gale, 1990) would be needed to decide whether such securities could be issued. However, the empirical evidence suggests that innovative financing strategies, with existing or synthetic securities, are worth exploring.

V. Conclusions

The optimal structure of government debt has been analyzed in a stochastic environment. In a setting with distortionary taxes, the government should smooth tax rates over states of nature as well as over time. This requires state-contingent government liabilities that provide a hedge against shocks to the budget.

For postwar U.S. data, tax smoothing as a positive theory of policy cannot be rejected on the basis of the time path of taxes. However, a number of security returns are correlated with tax rates, leading to a rejection on that basis. Estimates of optimal debt portfolios provide strong support for using nominal, nonindexed, government debt, but provide only weak evidence on the maturity distribution. Moreover, it seems that the government could improve tax smoothing by having some nontraditional liabilities, like foreign-currency debt or a short position in the stock market.

REFERENCES

- Allen, Franklin and Gale, Douglas, "Optimal Security Design," *Review of Financial Studies*, Fall 1988, 1, 229–63.
 Amemiya, Takeshi, *Advanced Econometrics*, Cambridge, MA: Harvard University Press, 1985.

- Anderson, Theodore W., *An Introduction to Multivariate Statistical Analysis*, New York: Wiley, 1958.
- Barro, Robert J., "On the Determination of Public Debt," *Journal of Political Economy*, October 1979, 87, 940-71.
- , "On the Predictability of Tax-Rate Changes," manuscript, University of Rochester and NBER, October 1981.
- Blanchard, Olivier and Fischer, Stanley, *Lectures on Macroeconomics*, Cambridge MA: MIT Press, 1989.
- Bohn, Henning, "Why Do We Have Nominal Government Debt?" *Journal of Monetary Economics*, January 1988, 21, 127-40.
- Calvo, Guillermo, "On the Time Consistency of Optimal Policy in a Monetary Economy," *Econometrica*, November 1978, 46, 1411-28.
- Fischer, Stanley, "Welfare Effects of Government Issue of Indexed Bonds," in Rudiger Dornbusch and Mario Simonsen, eds., *Inflation, Debt, and Indexation*, Cambridge, MA: MIT Press, 1983.
- Gale, Douglas, "The Efficient Design of Public Debt," in R. Dornbusch and M. Draghi, eds., *Capital Markets and Debt Management*, Cambridge: Cambridge University Press, forthcoming 1990.
- Hansen, Lars and Singleton, Kenneth, "Stochastic Consumption, Risk Aversion, and the Temporal Behavior of Asset Returns," *Journal of Political Economy*, April 1983, 91, 249-65.
- Judd, Kenneth, "Optimal Taxation in Dynamic Stochastic Economies: Theory and Evidence," manuscript, Hoover Institution, Stanford University, 1989.
- Kydland, Finn and Prescott, Edward, "Rules Rather Than Discretion: The Inconsistency of Optimal Plans," *Journal of Political Economy*, June 1977, 85, 473-91.
- Lucas, Robert E., "Asset Prices in an Exchange Economy," *Econometrica*, November 1978, 46, 1429-45.
- , and Stokey, Nancy, "Optimal Fiscal and Monetary Policy in an Economy without Capital," *Journal of Monetary Economics*, July 1983, 12, 55-93.
- Mehra, Rajnish and Prescott, Edward, "The Equity Premium: A Puzzle," *Journal of Monetary Economics*, March 1985, 15, 145-62.
- Modigliani, Franco and Sutch, Richard, "Innovations in Interest Rates Policy," *American Economic Review*, May 1966, 56, 178-97.
- Sahasakul, Chaipat, "The U.S. Evidence of Optimal Taxation Over Time," *Journal of Monetary Economics*, November 1986, 18, 251-75.
- Schmidt, Peter, *Econometrics*, New York: Marcel Dekker, 1976.
- Tresch, Richard, *Public Finance: A Normative Theory*, Plano, TX: Business Publications, 1981.

Beneficial Concentration

By ANDREW F. DAUGHETY*

Concentration is held by many economists (and undoubtedly by a large proportion of the population of economists) to be so worrisome as to be a cause for concern or even action. This concern is based on an intuition derived mainly from considering the symmetric equilibria of models of multi-firm industry behavior. Typically, as the number of firms in the symmetric equilibrium increases, some measure of welfare rises (e.g., surplus), and some measure of concentration falls (e.g., the Herfindahl index). However, there are two general "causes" of concentration in industries (too few firms and significant disparities in firm sizes), and the foregoing intuition does not carry over to the asymmetric equilibria that often lurk about in well-formulated models and which seemingly reflect a more realistic picture of the world. In this paper, I show that social optimality may involve extensive asymmetry: uniformity of firm size may be cause for concern.

This result leads to a number of implications. For example, concentration measures (such as the Herfindahl index) provide little insight about welfare, mainly because (as will be shown) increases in such measures will sometimes reflect increases in welfare and sometimes reflect decreases in welfare. A related implication will hold for average firm profits and concentration measures. In fact, as will be seen below, the often observed empirical result of a positive correlation between various measures of concentration and firm profits may indicate too

little asymmetry rather than too much. Finally, a merger that reduces the number of firms can be welfare-enhancing even if there are no cost advantages to the merger itself.

The analysis proceeds from a model that, in a very simple manner, varies the degree of asymmetry of the organization of the industry. Explaining why one asymmetric solution might be more reasonable than another is not of concern here. Rather, I will simply assume that such equilibria arise because it is individually (and possibly mutually) advantageous to the firms in the industry for this to occur. This paper focuses on the social consequences of such outcomes and the implications for policies associated with asymmetry (such as antitrust).

I. An Asymmetric Industry Comprised of Identical Firms

Consider an industry comprised entirely of identical firms. The reason for doing this is twofold. First, by eliminating interfirm differences, one can focus on attributes of the equilibrium, rather than characteristics of the technology or markets involved. Second, one eliminates any traditional social advantages that might accrue to asymmetric access to resource markets, technology, siting, or other factors.

Usually, when an industry comprised of identical firms is considered, one examines (i.e., restricts attention to) symmetric equilibria. This seemingly natural assumption ignores the fact that there are conditions under which asymmetric behavior can be individually and mutually beneficial to firms that are otherwise identical. Thus, for example, if one firm develops an R&D facility or develops a sophisticated marketing group, it is not always advantageous for the other firms in the industry to do likewise. Much depends upon the behavior of the potential leader, for by readily licensing new products, or by focusing the marketing cam-

*Departments of Economics and Management Sciences, College of Business Administration, The University of Iowa. This work was supported by NSF grant No. IST-8610360 and by a grant from the College of Business Administration, The University of Iowa. I thank George Neumann, Jennifer Reinganum, Mike Scherer, Gene Savin, Cliff Winston, and an anonymous referee for their suggestions and comments on earlier versions of this paper.

paigms on interindustry rather than intra-industry competition, the residual firms may choose to follow rather than also to invest in such types of capital. This occurs because firms in the real world have the flexibility to write the rules of the game that they wish to play.

Such asymmetry can be readily justified theoretically. For example, simply by allowing for production over time, various asymmetric equilibria may arise. Garth Saloner (1987) demonstrates this by allowing for two production periods before the market clears in the standard Cournot duopoly analysis with identical firms. He shows that all output combinations on the outer envelope of the best response functions from one Stackelberg solution to the other are sustainable as subgame perfect Nash equilibria. In this paper, I take as given that firms may find asymmetry advantageous and proceed to employ a model of identical firms that allows, in a very simple manner, for both symmetric and asymmetric equilibria. Specifically, a simple n -firm, static, quantity-setting oligopoly model, with the number of leaders as a parameter, will be posed and examined. Welfare (in the case at hand, measured by aggregate output), average profits, and concentration (measured by the Herfindahl index) can thus all be expressed as functions of the number of firms and the number of leaders, allowing for comparisons of interest.

To formalize the above, I first specify a static model of an industry comprised of n identical firms, each producing a homogeneous product at a constant marginal cost of c ; firm i 's output is denoted x_i . For convenience, let the inverse demand function which specifies price as a function of aggregate output be $p = a - b\sum_i x_i$, where $a > 0$, $b > 0$. Firm i 's profit is $\Pi^i(\mathbf{x}) = (a - c - b\sum_j x_j)x_i$, where $\mathbf{x} = [x_1, \dots, x_n]'$ is the vector of firm outputs.

A cautionary note is called for at this point. The foregoing setup is clearly quite special, especially with respect to the cost structure, since average costs are constant. It is straightforward, but notationally and algebraically unpleasant, to allow for declining average costs by (say) incorporating a

fixed cost in the cost function. If such costs are large, then the adjustments to industry structure that will be examined might induce exit, and thus the results would change. Technically, this is due to discontinuities in the optimal response functions for the firms (see Avinash Dixit [1979] for a discussion of this point). For moderate-to-small costs, no results will change. Since the main issue is to examine how industry structure can affect standard notions of the relationship between concentration and welfare, the above model (even though it is not particularly realistic) is sufficient. Moreover, since employing constant average costs is likely to underestimate (vis-à-vis decreasing costs) the welfare benefits of asymmetry, the analysis to follow errs, if at all, by being too conservative.

Since the main interest of this paper concerns industries of moderate size, I will not consider either collusive or perfectly competitive equilibria. Traditionally, two noncooperative oligopoly solutions have been employed to predict equilibrium output levels, namely Cournot and Stackelberg (see James W. Friedman [1977] and Daughety [1988] for detailed discussions of these solutions). An equilibrium is a Cournot (Nash) equilibrium if: 1) each firm chooses its output to be a best response (i.e., profit-maximizing response) to conjectured outputs of all other firms; and 2) the conjectures are correct for all firms. In the case of identical firms, the Cournot equilibrium is symmetric: all firms produce the same level of output and receive the same profit.

In the Stackelberg equilibrium, one firm is a "leader" and $n - 1$ firms are followers. The followers "play Cournot" by computing best-response output levels to the aggregate of all others' levels. The leader recognizes this and uses the followers' best-response functions to decide on a profit-maximizing output level. Thus, the leader's output level is not a best response to the followers' output levels; it is instead a best response to the followers' *best-response output functions*.

I will now extend this to allow for m leaders and $n - m$ followers (see Daughety, 1984; Hanif D. Sherali, 1984). Let the $n - m$ followers "play Cournot," taking the aggre-

gate of all other followers' output and the m leaders' output as given. Moreover, let the leaders recognize this (i.e., use the followers' best-response functions), but let each leader play Cournot against each other leader, realizing that all leaders understand this and are also choosing output in a similar manner. As will be seen below, this parameterization provides for a wide range of potential market structures.

Let x^L denote a typical leader's output, X^L denote aggregate leader output, x^F denote a typical follower's output, X^F denote aggregate follower output, and X^T denote aggregate industry output. Then, the m -leader equilibrium solutions are as follows:

$$x^L(m, n) = [(a - c)/b](m + 1)$$

$$X^L(m, n) = m[(a - c)/b]/(m + 1)$$

$$x^F(m, n) = [1/(n - m + 1)]x^L$$

$$= \frac{(a - c)/b}{(m + 1)(n - m + 1)}$$

$$X^F(m, n) = \frac{(n - m)[(a - c)/b]}{(m + 1)(n - m + 1)}$$

$$X^T(m, n) = [(n + m - m^2)/(n - m + 1)] \\ \times [(a - c)/b]/(m + 1).$$

Note that both $m = 0$ and $m = n$ correspond to a Cournot industry: when $m = 0$, all firms are followers and play Cournot; when $m = n$, all firms are leaders and play Cournot. Clearly, $m = 1$ corresponds to the standard Stackelberg model. It is tedious but straightforward to show the following (proofs are available from the author upon request).¹

- (1) For fixed n , X^T is concave in m ($\partial^2 X^T / \partial m^2 < 0$).
- (2) For fixed n , average firm profit, $\bar{\Pi}(m, n)$ ($\equiv \sum_i \Pi^i / n$), is convex in m .

Both results are intuitively reasonable. X^T concave in m extends the standard result that the Stackelberg aggregate output is greater than the Cournot aggregate output. Since $X^T(0, n) = X^T(n, n)$ and $X^T(1, n) > X^T(0, n)$, one would expect X^T to rise and then fall. This would also suggest that $\bar{\Pi}(m, n)$ would first fall and then rise. These two results figure significantly in what follows.

A. Welfare, Concentration, and Asymmetry

Under the assumptions used above, increases in aggregate output, X^T , result in increases in welfare (measured by surplus), and decreases in aggregate output result in decreases in welfare. Thus, welfare attains an interior maximum when X^T does, which is when the optimal² number of leaders, m^* , equals $n/2$.

Most importantly, welfare is maximized when there is considerable asymmetry: symmetric equilibria ($m = 0$ or $m = n$) are welfare-minimizing (for fixed n). In fact, for the type of asymmetry I am considering, any degree of asymmetry is socially preferable to symmetry. Again, this follows from the fact that X^T is concave in m and is symmetric about m^* .

One immediate further implication is that measures of concentration may have little, if anything, to do with indicating welfare. This is particularly easy to see with the Herfindahl index. The Herfindahl index is the sum of the squares of market shares of firms in the industry. George J. Stigler (1964) provides a theoretical basis for its use as a measure of concentration. To examine this

¹For expository convenience, m and n are being manipulated as if they were continuous variables. This is valid as a procedure, as long as appropriate rounding is employed as necessary. This holds since all the functions are either monotonic or unimodal, and thus (in one dimension) integer solutions can be directly constructed from the continuous solution.

²More precisely $m^* = n/2$ when n is even. When n is odd m^* is either $(n - 1)/2$ or $(n + 1)/2$. For example, if $n = 10$, then $m^* = 5$, while if $n = 11$, then, $m^* = 5$ or 6. Also, the use of X^T to measure welfare assumes that there are no *ex ante* differences between the equilibria for, say, m' and $m'' > m'$.

index in the setting herein, let S_L denote the aggregate output share of leader firms. Then, the Herfindahl index, H , is

$$H = \begin{cases} S_L^2/m + (1 - S_L)^2/(n - m) & \text{for } 1 < m < n - 1 \\ 1/n & \text{for } m = 0 \text{ or } n. \end{cases}$$

Again, treating m as a continuous variable, it can be shown that

$$\partial H / \partial m \begin{cases} \geq 0 & m \leq 1 \\ < 0 & m > 1. \end{cases}$$

Thus, H rises from $m = 0$ to $m = 1$ and then declines as m grows. While the algebra for this result is very messy, the intuition is reasonably straightforward. Clearly H is equal for $m = 0$ and for $m = n$, since these values produce symmetric industries; moreover, one would therefore expect H to rise and then fall as m ranges from zero to n , since symmetry minimizes H (for fixed n). In the situation with precisely one leader, there is one firm that has considerably greater share than any other firm in the industry. Additional leaders means that the aggregate leaders' share rises, but each individual leader's share does too. Therefore, when $m \geq 2$, there is no one firm with greater share than any other firm. Thus, one would expect H to peak at $m = 1$, which it does.

In general, therefore, H is declining both over ranges of m when welfare is increasing and over ranges of m when welfare is decreasing: decreasing concentration does not imply increasing welfare. A similar result obtains if a concentration ratio measure is used. For example, if the four-firm ratio is used, then this ratio increases as m increases up to $m = 4$, and then it falls monotonically as m continues to increase. This suggests that, in and of themselves, such measures do not reveal what one would most like to know: when competition has increased and welfare has improved.³

B. Profits and Concentration

In study after study, a positive correlation between average firm profit and various measures of concentration has been found (see Leonard W. Weiss, 1974; Frederick M. Scherer, 1980). What this correlation means is less obvious (e.g., Sam Peltzman, 1977; Scherer, 1980). The foregoing analysis indicates that profits and concentration measure are positively correlated when the number of leaders is less than m^* , the socially optimal number. To see this, recall that $\bar{\Pi}(m, n)$ is average firm profit. This can be readily shown to be

$$\bar{\Pi}(m, n) = \frac{(a - c)^2(n + nm - m^2)}{nb(n - m + 1)(m + 1)}.$$

As indicated earlier, $\bar{\Pi}(m, n)$ is convex in m , and its minimum occurs at $m^* = n/2$. Therefore, for $m^* > 1$, $\bar{\Pi}(m, n)$ declines while H declines for $m < m^*$, and $\bar{\Pi}(m, n)$ rises as H declines for $m > m^*$. In other words, for typical industries ($n > 2$), there is a positive correlation between average firm profits and concentration when there are too few leaders and a negative correlation when there are too many. Figure 1 illustrates the relationships among X^T , $\bar{\Pi}$, and H .

Thus, the empirical observation that there is a robust positive correlation between average firm profits and industry concentration may simply mean that the ratio of leaders to firms in the typical industry in the sample is low (i.e., not socially optimal), since actual m appears to be to the left of m^* .

C. Mergers and Welfare

Up to this point, n has been held fixed. While large changes in n (e.g., n becoming infinite) have the usual results, it is small

an index, however, cannot rely upon information on market share alone, as is evident from the foregoing material and from the discussion of the industry performance gradient index in Robert E. Dansby and Robert D. Willig (1979).

³This is not to say that one could not construct an index that would reflect welfare considerations. Such

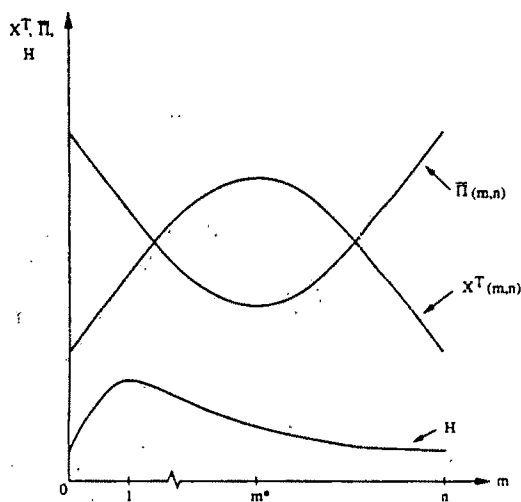


FIGURE 1. WELFARE, AVERAGE PROFITS, AND CONCENTRATION

changes in n that are now of interest. For example, what if two firms merge? As will be shown below, this need not mean that welfare has fallen. As pointed out earlier, there is no cost advantage to a merger here. Thus, any advantages or disadvantages are purely due to the noncooperative equilibrium itself.

There are several recent papers that have examined the strategic effects of mergers. Stephen W. Salant et al. (1983) examine a quantity-setting oligopoly and look at the conditions under which merger may be disadvantageous to the participants. This can occur since, by merging and restricting output, the residual firms in the industry can expand, with the net result being that the merger is disadvantageous for the participants. Raymond Deneckere and Carl Davidson (1985) demonstrate how this result can be sensitive to the assumption of strategic variable employed, by employing a model in prices and getting opposite results. Morton Kamien and Israel Zang (1987) employ a model of a suboptimizing holding company in a quantity-choosing context that reestablishes a variety of conditions wherein merger can be advantageous. Most recently, Joseph Farrell and Carl Shapiro (1990) use a Cournot model to analyze mergers and also observe that mergers can be welfare-enhancing. They find that if the merger

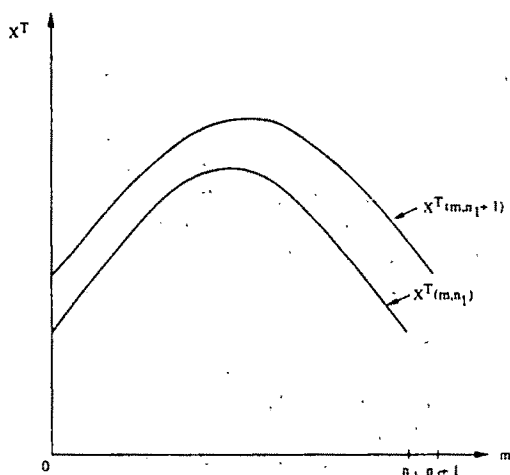


FIGURE 2. WELFARE EFFECTS OF CHANGES IN m AND n

generates "synergies" (e.g., if the participants in the merger can recombine their assets so as to increase their joint production possibilities), then welfare can be enhanced by a merger. As will be seen below, welfare can be enhanced by a merger in spite of the lack of such synergies, if the merger alters the behavior of the participants.

In this section, I wish to examine a special type of merger (or dismemberment) wherein, say, two followers merge and the result is a firm that *behaviorally* is a leader. Figure 2 illustrates $X^T(m, n)$ for industries with $n = n_1$ and $n = n_1 + 1$ (say 12 and 13) firms. Recalling that equilibrium total output for an n -firm industry with m leaders is

$$X^T(m, n) = \frac{(a - c)(n + nm - m^2)}{(m + 1)(n - m + 1)b}$$

it is straightforward to show that

$$\partial X^T / \partial m \begin{cases} < 0 & n < 2m \\ = 0 & n = 2m \\ > 0 & n > 2m \end{cases}$$

$$\partial X^T / \partial n > 0$$

$$\partial^2 X^T / \partial m \partial n \begin{cases} < 0 & 3m + 1 < n \\ = 0 & 3m + 1 = n \\ > 0 & 3m + 1 > n. \end{cases}$$

As indicated in the figure, changes in n alone have the usual interpretation. Thus, for example, a bankruptcy accompanied by no entry and no change in industry structure is welfare-impairing. The cross-partial, however, is most interesting: could changes in industry structure result in a welfare improvement? What if a merger reduced the number of firms but increased m ? Clearly, this is a very special type of merger, such as was briefly discussed above.

More precisely when is $X^T(m+1, n-1)$ greater than $X^T(m, n)$? A little algebraic manipulation reveals that this occurs when $3(m+1) < n$. Thus, leader-generating mergers in industries that are very close to symmetric (and low in terms of the number of leaders) *can* be socially desirable.⁴ The increase in asymmetry again results in an increase in competition, resulting in greater aggregate output and greater welfare. Of course, as the derivative with respect to n shows, mergers that simply reduce the number of followers without increasing the number of leaders reduce welfare.

Of course, such mergers should not occur unless the new entity is more profitable than the individual parts. Thus, when is $\Pi^L(m+1, n-1) > 2\Pi^F(m, n)$ (i.e., when does the postmerger profit of the generated leader exceed the premerger profits of the two followers)? This turns out to occur if and only if $f(m, n) \equiv (n-m+1)^2(m+1)^2 - 2(n-m-1)(m+2)^2 > 0$ for the premerger industry with n firms and m leaders, a very unpleasant condition. However, a grid search over all possible values of n and m such that $3(m+1) < n$ yields a very nice result: for $n \leq 28$, if a leader-generating merger is welfare-enhancing, it is profit-enhancing!⁵ This contrasts with the existing literature on merger analyses performed with static quantity models; Salant et al. (1983) show conditions wherein using a Cournot oligopoly model to examine merg-

ers can imply profit losses from mergers for the participants, due to equilibrium output adjustments. In general, the size of merger that is considered here would show losses under a Cournot model.

Thus, when m is small compared to n (i.e., somewhat less than one-third of the firms are leaders), mergers that are leader-generating can be both privately and socially advantageous. However, this is not to say that all mergers that increase the number of leaders are welfare-enhancing (or that all divestments are welfare-impairing⁶). As indicated earlier, too many chefs can spoil the broth: $\partial X^T / \partial m$ is negative when $m > n/2$. In this case (in equilibrium), dismemberment of a firm which would decrease the stock of leaders and increase the stock of followers would be socially beneficial. The main point is, however, that mergers can be welfare-enhancing, even without cost advantages or other traditional considerations.

II. Conclusions

The point of this paper is very simple: asymmetry, and thus concentration, can be beneficial from society's viewpoint. This benefit does not depend upon scale economies or marketing advantages or learning-by-doing. It derives entirely from the noncooperative nature of firm interaction. Alternatively put, asymmetric organization can be socially as well as individually optimal. Thus, focusing on concentration as deleterious is misguided: actions that reduce the number of firms or increase concentration can be welfare-enhancing.

The fundamental attribute that contributed to this result was rational asymmetry in firm strategic behavior. Firms adopt roles in industries; as indicated earlier, not all firms in an industry find it beneficial to operate market-forecasting groups or to en-

⁴Alternatively, when $3m-1 < n$, then $X^T(m, n) > X^T(m-1, n+1)$; that is, breaking up a leader firm into two followers reduces welfare, even though the total number of firms has increased.

⁵When $n > 28$, the quartic properties of $f(m, n)$ make characterization difficult.

⁶Welfare-enhancing dismemberment of a leader that produces two followers simply reverses the stated condition. From the derivative condition, mergers of two followers into a follower always lower welfare. Since $X^T(m, n) > X^T(m-1, n-1)$, merger of two leaders into a leader always lowers welfare.

gage in R&D, especially if other firms in the industry do such things. The Nash equilibria that support such outcomes can involve greater aggregate output than those associated with symmetric roles. That such individually rational behavior can be socially advantageous is simply one more reason for avoiding single-minded adherence to the siren song of symmetry, both in analysis and (especially) in policy.

REFERENCES

- Dansby, Robert E. and Willig, Robert D., "Industry Performance Gradient Indexes," *American Economic Review*, June 1979, 69, 249-60.
- Daughety, Andrew F., "Endogenously Determined Industrial Organization," Working Paper 84-34, College of Business Administration, University of Iowa, November 1984 (revised March 1985).
- _____, *Cournot Oligopoly: Characterization and Applications*, Cambridge: Cambridge University Press, 1988.
- Deneckere, Raymond and Davidson, Carl, "Incentives to Form Coalitions with Bertrand Competition," *Rand Journal of Economics*, Winter 1985, 16, 473-86.
- Dixit, Avinash, "A Model of Duopoly Suggesting a Theory of Entry Barriers," *Bell Journal of Economics*, Spring 1979, 10, 20-32.
- Farrell, Joseph and Shapiro, Carl, "Horizontal Mergers: An Equilibrium Analysis," *American Economic Review*, March 1990, 80, 107-26.
- Friedman, James W., *Oligopoly and the Theory of Games*. Amsterdam: North Holland, 1977.
- Kamien, Morton and Zang, Israel, "The Limits of Monopolization Through Acquisition," Discussion Paper No. 754, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, 1987.
- Peltzman, Sam, "The Gains and Losses from Industrial Concentration," *Journal of Law and Economics*, October 1977, 20, 229-63.
- Salant, Stephen W., Switzer, Sheldon and Reynolds, Robert J., "Losses from Horizontal Merger: The Effect of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium," *Quarterly Journal of Economics*, May 1983, 98, 185-99.
- Saloner, Garth, "Cournot Duopoly with Two Production Periods," *Journal of Economic Theory*, June 1987, 42, 183-7.
- Scherer, Frederick M., *Industrial Market Structure and Economic Performance*, 2nd ed., Chicago: Rand McNally, 1980.
- Sherali, Hanif D., "A Multiple Leader Stackelberg Model and Analysis," *Operations Research*, March/April 1984, 32, 390-404.
- Stigler, George J., "A Theory of Oligopoly," *Journal of Political Economy*, February 1964, 72, 44-61.
- Weiss, Leonard E., "The Concentration-Profits Relationship and Antitrust," in H. J. Goldschmid, H. M. Mann, and J. F. Weston, eds., *Industrial Concentration: The New Learning*. Boston: Little Brown, 1974, 201-20.

Horizontal Mergers: The 50-Percent Benchmark

By DAN LEVIN*

The paradigm often used in the study of industrial organization relates market structure to conduct and performance (e.g., F. M. Scherer, 1980 pp. 4-5). Measures such as concentration ratios and the Herfindahl-Hirschman index provide partial characterization of structure by capturing the number and size distribution of firms.

Horizontal merger among a subset of firms in the same market may reduce competition by reducing the number of firms and increasing concentration. If the merged firm can lower costs by reallocating production, the incentive to merge is reinforced. The concern is that such a change in structure increases market power and adversely affects market performance. Many economists dismiss or are skeptical about favorable effects of horizontal mergers. Scherer (1980 p. 546) concludes: "an impressive accumulation of evidence points to the conclusion that mergers seldom yield substantial cost savings, real or pecuniary."

Such views are the main rationale for antitrust policy in this area. Section 7 of the Clayton Act was designed to thwart monopoly power "in its incipency"; it speaks in terms of a merger that "may be substantially to lessen competition or to tend to create a monopoly." Yet, antitrust policy toward horizontal mergers is evolving. The Department of Justice adopted new guidelines in 1982 with modest changes in 1984. Further proposals were made during the Reagan Administration.

The four papers in the recent Symposium on "Horizontal Mergers and Antitrust" in *The Journal of Economic Perspectives* (Fall 1987), clearly demonstrate the complexity of

the issues involved in horizontal mergers: market definition, market concentration, ease of entry, and efficiencies. I abstract here from issues of market definition or entry and focus on the effect such mergers have on performance and welfare vis-à-vis their impact on concentration, market power, and conduct. In other words, this paper contributes to the debate described so nicely by Lawrence White (1987 p. 14), who poses it in terms of the following two hypotheses: is it the case that "the more easily a group of sellers (who collectively might have market power) can coordinate and police their mutual actions, the more likely are they to approximate a monopoly outcome"; or is it the case that "it only takes two to make a horse race"?

I analyze the consequences of a horizontal merger by a subset of firms in an industry, assuming that firms that are not part of the merger behave à la Cournot. My analysis is related to that of Stephen Salant, Sheldon Switzer, and Robert Reynolds (1983) (SSR hereafter), who study the consequences of a horizontal merger by a subset of firms in a Cournot industry. Unlike SSR, I do not restrict the merged group to remain a Cournot player after the merger.¹ Thus, the merged group can become a Stackelberg leader, a conjectural variation player, or remain Cournot.² I impose only

¹SSR assume that the merged group behaves like a multiplant player who engages in a noncooperative game against other firms after the merger. In the case of symmetric, identical constant marginal cost, this assumption implies that, no matter how many firms merge, their total postmerger output will be the same as that of one single firm that stays out. I find this implication to be restrictive.

²The literature in industrial organization recognizes the possible need to assign different perceptions and modes of behavior to separate firms whose sizes or cost structures are asymmetrical. See Scherer (1980 pp. 176, 232-33) for a discussion of the dominant firm model and Hal Varian (1984 pp. 101-3) for a discussion of

*Department of Economics, University of Houston, Houston, TX 77204-5882. The financial support of the Center for Public Policy at the University of Houston is gratefully acknowledged. I thank, without implicating, Franklin Fisher, my colleague James Smith, and an anonymous referee for comments and suggestions.

stability conditions and proceed to study the implications of such horizontal mergers on profits and welfare.³

I find that, if a group of firms with less than 50 percent of market output considers a horizontal merger then a) any contraction of output by the merged group will cut profits below the level obtained by only real-locating their premerger output⁴ and b) any profitable merger will raise welfare. These results may have strong implications for public policy in this area. With market boundaries defined:

The [new (1982)] Guidelines use the Herfindahl-Hirschman Index (HHI) as their primary market concentration guide, with concentration levels of 1000 and 1800 as their two key levels. Any merger in a market with a postmerger HHI below 1000 is unlikely to be challenged; a merger in a market with a postmerger HHI above 1800 is likely to be challenged (if the merger partners have market shares that cause the HHI to increase by more than 100), unless other mitigating circumstances exist, like easy entry. Mergers in markets with post-concentration HHI levels between 1000 and 1800 require further analysis before a decision is made whether to challenge" [White, 1987 p. 16].

Taken at face value, my analysis suggests that a two-firm merger should not be challenged when their premerger market share

is less than 50 percent, which is assured when the HHI is less than 1250. Some limitations of the analysis are discussed in the summary section.

I. The Basic Model

Let there be $n > 1$ Cournot firms that sell their homogeneous output at the market price, P . The market (inverse) demand function of total industry output, Q , is $P(Q)$ with $P' < 0$ where $P > 0$ and is assumed to be twice continuously differentiable. I also assume the following.

ASSUMPTION 1 (A1): *Firms have cost functions of the form $C_i(q_i) = c_i q_i$, where $q_i \geq 0$ is the output of the i th firm.*

Let $c = \min_i(c_i)$. Clearly with A1 we must have $P(0) > c$ in order for firms ever to produce positive outputs. Define output level \underline{Q} by $P(\underline{Q}) = c$.

Different marginal cost provides additional incentive to merge. To focus on incentives for horizontal merger that relate strictly to market power, concentration, and conduct, it will sometimes be useful to abstract from efficiency considerations and assume, as in SSR, the following.

ASSUMPTION 1* (A1*): *All firms have identical cost functions, $C(q_i) = cq_i$.*

Next, I impose the following restrictions on the demand function.

ASSUMPTION 2 (A2): $P'(Q) + P''(Q)q_i < 0$ for all $0 \leq q_i \leq Q$, as long as $P > 0$.

Frank Hahn (1962) makes and defends A2. It requires that, at all possible outputs, the marginal revenue of any one producer with a given output is a diminishing function of total output of his rivals. Roy Ruffin (1971) assumes A2, demonstrates its reasonableness, and suggests an alternative interpretation, namely that A2 requires that, at all possible outputs, the marginal revenue function facing any firm is steeper than the demand function. A2, together with another assumption that is trivially satisfied by A1,

the Stackelberg leader and conjectural-variations models. An overview of recent developments in the theory of horizontal mergers is in Gerard Gaudet and Salant (1989).

³Welfare analysis is absent from SSR's work.

⁴SSR show that it is possible and even plausible that such a horizontal merger, within a Cournot industry, is not profitable. Their linear-demand example shows that, if the subset of firms that merged has less than 80 percent of premerger output, then their total profits will be lower. One may ask how sensitive this market-share mark is to the linear-demand and identical-marginal-cost assumptions. If by dropping linearity a small percentage of market output by the merged group may assure an increase in the group profits and reduction in welfare, then the SSR demonstration, though interesting, has little bearing on policy issues.

has been shown to assure the stability of Cournot oligopoly.⁵ At the very least, A2 is an important extension of the example in SSR which assumes that P'' is identically zero.

It is well known that A1 and A2 are sufficient for the existence of a unique Cournot-Nash Equilibrium (CNE).⁶ The unique CNE is completely characterized by the first-order conditions for profit maximization of n firms:

$$(1) \quad P(Q) + P'(Q)q_i - c_i = 0, \quad i = 1, \dots, n.$$

Summing over i and rearranging yields

$$(2) \quad P(Q) + \frac{P'(Q)Q}{n} - \sum_{i=1}^n \frac{c_i}{n} = 0.$$

Denote by Q^* total industry output at the CNE with n firms, by $q_i^* > 0$ each firm's output, by $P^* = P(Q^*)$ the market price, and by $\pi_i^* = q_i^*(P^* - c_i)$ each firm's profits.

II. Horizontal Mergers: Incentives and Welfare Implications

A subset of m firms, $2 \leq m \leq n$, is considering a merger. Denote this set of firms by M , their total output by Q_M , and their premerger total output by $Q_M^* = \sum_{j \in M} q_j^*$. Denote by F the fringe of $k \equiv n - m$ firms that stay out of the merger and by Q_F , F 's total output.

ASSUMPTION 3 (A3): *The k members of F remain Cournot firms after a merger.*

This assumption is made in SSR, and though it is potentially restrictive, I maintain it here

⁵Hahn (1962) shows that A2 together with the assumption $P' - C'' < 0$ for all firms assures stability of Cournot oligopoly. A recent work by A. Al-Nowaihi and P. L. Levine (1985) shows that Hahn's conditions assure only local stability. It is global only if the number of firms is less than five.

⁶Levin (1982) provides a direct proof that allows any continuous and twice-differentiable $C_i(q_i)$, as long as $P' - C''(q_i) < 0$, which is trivially satisfied by A1. This result can be shown to be a special case of a much more general proof in B. Rosen (1965).

to simplify the analysis. However, unlike SSR, which requires that M remain a Cournot firm as well, I impose no restrictions except cost minimization on M 's mode of behavior. In other words, though I do not model the mechanism that determines M 's behavior after the merger, I do allow the possibility that M may become a Stackelberg leader, a "conjectural variation" firm, stay Cournot as in SSR, or anything else.

Clearly $Q_M < \underline{Q}$. Under A1, A2, and A3, Levin (1982) shows that for any given $Q_M < \underline{Q}$ there is a unique CNE for firms in F satisfying for all $i \in F$

$$(3) \quad P(Q_M + Q_F) + P'(Q_M + Q_F)q_i - c_i \leq 0$$

with strict inequality only if $q_i = 0$. Denote by $q_i(Q_M)$ and $Q_F(Q_M)$ the output of the i th firm in F and the total output of F when the output of M is Q_M .

LEMMA 1: *Any $0 \leq Q_M^0 \leq Q_M \leq \underline{Q}$ implies $Q_M^0 + Q_F(Q_M^0) \leq Q_M + Q_F(Q_M)$ and $q_i(Q_M^0) \geq q_i(Q_M)$ for all $i \in F$.*

PROOF:

Suppose $Q_M^0 \leq Q_M$ but that $Q_M^0 + Q_F(Q_M^0) > Q_M + Q_F(Q_M)$. This implies, due to A2 and (3), that $P(Q_M^0 + Q_F(Q_M^0)) + P'(Q_M^0 + Q_F(Q_M^0))q_i(Q_M^0) - c_i < 0$ for all $i \in F$. Since $P' < 0$, it implies that $q_i(Q_M^0) \leq q_i(Q_M)$ for all $i \in F$, so that $Q_F(Q_M^0) \leq Q_F(Q_M)$. Thus, $Q_M^0 + Q_F(Q_M^0) \leq Q_M + Q_F(Q_M)$, which leads to a contradiction. Hence $Q_M^0 \leq Q_M$ implies $Q_M^0 + Q_F(Q_M^0) \leq Q_M + Q_F(Q_M)$. The last inequality implies, due to A2 and (3), that for all $i \in F$, $P(Q_M + Q_F(Q_M)) + P'(Q_M + Q_F(Q_M)) \times q_i(Q_M^0) - c_i \leq 0$. Thus, $q_i(Q_M^0) \geq q_i(Q_M)$ for all $i \in F$, establishing the proof.

The uniqueness of the CNE for F implies that, if $Q_M = Q_M^*$, $q_i(Q_M^*) = q_i^*$ for all $i \in F$. Thus, in such a case, total industry output and price remain as in the premerger equilibrium.

Distinguish between $Q_M \leq Q_M^*$ and $Q_M > Q_M^*$, namely, M contracts or expands output relative to its premerger level.

Case 1: $Q_M \leq Q_M^*$ (contraction). From Lemma 1, for all $i \in F$, $q_i(Q_M) \geq q_i^* > 0$. Thus, as long as $Q_M \leq Q_M^*$, k firms with strictly positive outputs remain in F , so that (3) holds with equality. Summing over all $i \in F$ in (3) and rearranging yields

$$(4) \quad P(Q_M + Q_F) + \frac{P'(Q_M + Q_F)Q_F}{\bar{c}_F} - \bar{c}_F = 0$$

where $\bar{c}_F = \sum_{i \in F} c_i / k$, the average marginal cost of firms in F .

In the absence of fixed costs, (4) defines Q_F , and thus each q_i , $i \in F$, as a continuous function depending only on Q_M and \bar{c}_F but not on the distribution of the c_i 's. This property, discussed in Theodore Bergstrom and Hal Varian (1985a,b), is very useful. It implies (since \bar{c}_F remains constant) that all variables of the model can be expressed in terms of Q_M . Differentiating (4) with respect to Q_M yields

$$(5) \quad \beta \equiv \frac{d(Q_M + Q_F)}{dQ_M} = \frac{P'}{kP' + P' + P''Q_F}.$$

β is the response of total industry output to an increase in the output of M . Since $Q_F \leq Q$, $P' + P''Q_F \leq \max[P', P' + P''Q] < 0$, where the last inequality is due to A2. Thus, using (5), I conclude that

$$(6) \quad 0 \leq k\beta < 1.$$

Case 2: $Q_M > Q_M^*$ (expansion). From Lemma 1, for all $i \in F$, $q_i(Q_M) \leq q_i^*$, yet $Q_M + Q_F(Q_M) \geq Q_M^* + Q_F(Q_M^*)$; that is, total industry output increases when M expands its output.

A. Profitability of Horizontal Mergers

Define a merger as *profitable* for a subset m of firms if the postmerger profits of M exceed the sum of profits the m firms in M have without a merger. I consider first a contraction by the merged firms; that is, $Q_M \leq Q_M^*$ (Case 1). Let c_M be the lowest

marginal cost of any firm in M , this is $c_M = \min(c_j)$, $j \in M$, and let R be the cost saving to M as a result of producing the premerger output, Q_M^* , but in the lowest-marginal-cost plant(s). This is $R = \sum_{j \in M} (c_j - c_M)q_j^* \geq 0$.

THEOREM 1: Assume A1, A2, A3, and $c_M \leq \bar{c}_F$, and consider a merger by any subset m of firms that have no more than 50 percent of the premerger market output. Any contraction of output by the merged firm will reduce profit below the level obtained by simply real-locating their premerger output.

PROOF:

To minimize the cost of M , production must take place only at plant(s) with marginal cost c_M . Thus, actual profits for M are $\pi_M = Q_M[P(Q_M + Q_F(Q_M)) - c_M]$. $Q_M^* \leq Q_F^*$ (the 50-percent assumption) and $dQ_F/dQ_M = \beta - 1 < 0$ from (6). Hence, any $Q_M \leq Q_M^*$ implies $Q_F(Q_M) \geq Q_M^* \geq Q_M$. Then,

$$\begin{aligned} d\pi_M/dQ_M &= P(Q_M + Q_F(Q_M)) - \bar{c}_F + \bar{c}_F - c_M \\ &\quad + P'(Q_M + Q_F(Q_M))Q_M\beta \\ &= -P'(Q_M + Q_F(Q_M))Q_F(Q_M)/k \\ &\quad + P'(Q_M + Q_F(Q_M))Q_M\beta + \bar{c}_F - c_M \\ &= (-P'/k)[Q_F(Q_M) - Q_Mk\beta] + \bar{c}_F - c_M \\ &> (-P'/k)[Q_F(Q_M) - Q_M] + \bar{c}_F - c_M \\ &\geq \bar{c}_F - c_M \geq 0. \end{aligned}$$

The first equality is due to (5), the second equality is due to (4), the strict inequality is due to $k\beta < 1$ from (6) and the fact that $(-P'/k) > 0$, the first inequality is due to $Q_F(Q_M) - Q_M \geq 0$, and the last inequality is by assumption. Since $d\pi_M/dQ_M > 0$ at any $Q_M \leq Q_M^*$, $Q_M < Q_M^*$ implies that $\pi_M < \pi_M^* \equiv Q_M^*(P^* - c_M)$.⁷

⁷The result of Theorem 1 can be extended to cost functions of the form $C_i(q_i) = c_i q_i + (1/2)dq_i^2$ as long as $d \leq 0$ and the second stability assumption discussed in footnote 5 is satisfied ($p' - d < 0$).

It is possible that a merger such as in Theorem 1 is profitable with contraction when $R > 0$. However, contraction cannot be profitable if one eliminates reallocation of output as an incentive for such merger, as in the following corollary.

COROLLARY: Assume A1*, A2, and A3, and consider a merger as in Theorem 1. Any contraction of output by the merged firm will cut profit (i.e., $\pi_M < \pi_M^*$).

PROOF:

Here, $c_M = c = \bar{c}_F$, $R = 0$, and the rest of the proof mimics the proof to Theorem 1. One concludes that $Q_M < Q_M^*$ implies $\pi_M < \pi_M^*$.

B. Horizontal Mergers and Welfare

I use the sum of consumer and producer surplus, ignoring income effects, as a welfare measure denoted by W . Let $W(Q_M)$ be the postmerger level of welfare given that M produces Q_M , and let $\Delta W(Q_M \geq Q_M^0)$ measure the change in welfare as a result of M increasing its output from Q_M^0 to Q_M . This is

$$\begin{aligned} (7) \quad \Delta W(Q_M \geq Q_M^0) &= W(Q_M) - W(Q_M^0) \\ &= \int_{Q_M^0 + Q_F(Q_M^0)}^{Q_M + Q_F(Q_M)} P(t) dt \\ &\quad - \sum_{i \in F} c_i [q_i(Q_M) - q_i(Q_M^0)] \\ &\quad - c_M (Q_M - Q_M^0). \end{aligned}$$

Let c_F be the lowest marginal cost of any firm in F, this is $c_F = \min(c_i)$, $i \in F$.

THEOREM 2: Under A1, A2, A3, and $c_F \geq c_M$, as long as the profit of the merged firm remains positive, any increase in its output raises welfare.

PROOF:

By Lemma 1, $Q_M > Q_M^0$ implies $Q \equiv Q_M + Q_F(Q_M) \geq Q_M^0 + Q_F(Q_M^0) \equiv Q^0$. Therefore $\int_{Q^0}^Q P(t) dt \geq P(Q)(Q - Q^0)$. Since in this case $q_i(Q_M) \leq q_i(Q_M^0)$ for all $i \in F$, then $\sum_{i \in F} c_i [q_i(Q_M) - q_i(Q_M^0)] \leq c_F [Q_F(Q_M) - Q_F(Q_M^0)]$. Thus,

$$\begin{aligned} \Delta W(Q_M \geq Q_M^0) &\geq P(Q)(Q - Q^0) \\ &\quad - c_F [Q_F(Q_M) - Q_F(Q_M^0)] \\ &\quad - c_M (Q_M - Q_M^0) \\ &\geq P(Q)(Q - Q^0) - c_M (Q - Q^0) \\ &= (P(Q) - c_M)(Q - Q^0) \geq 0. \end{aligned}$$

The second inequality is due to $c_F \geq c_M$ and the fact that $Q_F(Q_M) - Q_F(Q_M^0) \leq 0$, and the last inequality is due to the fact that both terms in the product are nonnegative.

The intuition of this result is simple. When M increases its output, F contracts its output, but by a smaller amount. Hence, one can decompose such a change into two distinct parts: a pure reallocation of production from F to M and a pure increase in the total output of this industry. The net increase in output is welfare-enhancing, since price is above marginal cost. The reallocation of output can adversely affect welfare. The assumption, $c_F \geq c_M$, assures that the reallocation part is also welfare-enhancing.⁸ Often welfare increases with the output of M under weaker conditions. Note that this

⁸With linear demand, the weaker $\bar{c}_F \geq c_M$ assures a favorable reallocation affect as a result of an increase in Q_M . However, when Q_M increases beyond Q_M^* , the number of firms in F with positive output may fall, and \bar{c}_F will get smaller as a result of high-marginal-cost firms dropping out.

result is independent of the market share of M . Also note that under $A1^*$, $c_F = c_M$, so welfare increases with the output of M .

Let $\Delta W(Q_M)$ measure the change in welfare as a result of a merger M that produces Q_M . Clearly $\Delta W(Q_M) = W(Q_M) - W(Q_M^*) + R$. If the reallocation of output (efficiency) incentives for merger are removed (i.e., $R = 0$), one obtains the following

THEOREM 3: *Under $A1^*$, $A2$, and $A3$, any profitable merger by any subset m of firms that have no more than 50 percent of the premerger market output will raise welfare.*

PROOF:

Under these conditions, the corollary to Theorem 1 shows that a profitable merger implies $Q_M > Q_M^*$. This implies, by Theorem 2, that welfare will increase as long as $c_F - c_M \geq 0$; but under $A1^*$, $c_F = c = c_M$.

Could Theorem 3 be extended to the weaker $A1$ rather than $A1^*$? The point is that under $A1$ there are possible reallocation gains to M ($R > 0$), so it is possible to end up with profitable mergers where $Q_M < Q_M^*$. Are the efficiency gains, R , sufficient to compensate for such possible contraction in Q_M and in total output Q ? The answer is yes, if one assumes $P'' \leq 0$ rather than $A2$.

THEOREM 4: *Under $A1$, $P'' \leq 0$, $A3$, and $c_F - c_M \geq 0$, a profitable merger by any subset m of firms that have no more than 50 percent of the premerger market output will raise welfare.*

PROOF:

By Theorem 2, I need to show the above only for $Q_M^0 < Q_M^*$ (contraction) and, in light of Theorem 1, only for $R > 0$. Since $Q_M^0 < Q_M^*$, it is known that (4), (5), and (6) hold and that (3) holds with equality.

Let $\Delta f(Q_M^0) \equiv f(Q_M^0) - f(Q_M^*)$ and denote consumer surplus by CS . $\Delta W(Q_M^0) = \Delta CS(Q_M^0) + \Delta \pi_F(Q_M^0) + \Delta \pi_M(Q_M^0) \geq \Delta CS(Q_M^0) + \Delta \pi_F(Q_M^0)$, since $\Delta \pi_M(Q_M^0) \geq 0$ by presumption. $\Delta CS(Q_M^0) > -(P^0 - P^*)Q^*$ (i.e., the negative of the increase in market price times the larger quantity $Q^* > Q^0$).

Then,

$$\begin{aligned} \Delta \pi_F(Q_M^0) &= \sum_{i \in F} \int_{Q_M^*}^{Q_M^0} \pi_i'(t) dt \\ &= \sum_{i \in F} \int_{Q_M^*}^{Q_M^0} \{q_i' [P(t + Q_F(t)) - c_i] \\ &\quad + q_i(t) P'(t + Q_F(t)) \beta\} dt \\ &= \int_{Q_M^*}^{Q_M^0} P'(t + Q_F(t)) \beta \\ &\quad \times \left[Q_F(t) + \sum_{i \in F} (-q_i') q_i / \beta \right] dt \end{aligned}$$

by using (3) where $\pi_i' = d[q_i(Q_M)(P(Q_M + Q_F(Q_M)) - c_i)]/dQ_M$ and where $q_i' \equiv dq_i(Q_M)/dQ_M = -(P' + P''q_i)/(kP' + P' + P''Q_F) < 0$, for all $i \in F$, which is obtained by differentiating (3). However, under $P'' \leq 0$, $-q_i'$ and q_i are positively correlated; thus, $\sum_{i \in F} (-q_i') q_i / \beta \geq (Q_F/k)(1 - \beta)/\beta$ and also $(1 - \beta)/\beta \geq k$ [under $P'' \leq 0$, (5) implies that $0 < (k + 1)\beta < 1$], implying that $\sum_{i \in F} (-q_i') q_i / \beta \geq Q_F$. Thus,

$$\begin{aligned} \Delta \pi_F(Q_M^0) &\geq \int_{Q_M^*}^{Q_M^0} -P'(t + Q_F(t)) \beta 2Q_F(t) dt \\ &\geq Q^* \int_{Q_M^*}^{Q_M^0} -P'(t + Q_F(t)) \beta dt \\ &= Q^*(P^0 - P^*). \end{aligned}$$

The first inequality is due to $Q_M^* > Q_M^0$ and $-P' > 0$. The second inequality is due to $2Q_F(t) \geq 2Q_F^* \geq Q^*$ because of both the fact that Q_F expands as Q_M contracts and the 50-percent assumption. The conclusion is that $\Delta W(Q_M^0) > 0$. This establishes the proof.

Profitability of the merger is measured in terms of actual profits of M rather than perceived profits. If M becomes a Stackelberg leader, the difference disappears since such M correctly anticipates the response of

F.⁹ In other cases, unprofitable Q_M are not stable in the sense that M can dissolve itself or imitate what it did before the merger.

C. A Few Additional Observations

1. The 50-percent mark in the theorems is not necessary. Simple computations show that, if demand is linear and if marginal costs are the same, a merger of two firms in a three-firm market is either unprofitable or welfare-enhancing.

2. When M remains Cournot after the merger, it can be shown (using M's first-order condition for maximization and A2) that output will not increase (see Joseph Farrell and Carl Shapiro [1990] for this result with a more general cost structure). Thus, Theorem 2 will not apply, but Theorem 4 will: namely, if conditions of Theorem 4 are satisfied, profitable mergers are welfare-enhancing even if output falls.

3. Though $Q_M > Q_M^*$ implies $W > W^*$, not every expansion in Q_M is profitable. A very large Q_M , so that P is close enough to c_M , is clearly not profitable. However, a small expansion in Q_M will raise W and is profitable, as the derivative of the profit function in the proof to Theorem 1 suggests. Consider a case in which M becomes a Stackelberg leader that maximizes its profits while fully accounting for F's reaction. Since M may choose $Q_M = Q_M^*$, it must be able to increase its profit. The proof indicates that it must be with $Q_M > Q_M^*$.¹⁰

4. The proof in Theorem 1 relies on the condition that $k[1 + dQ_F/dQ_M] < 1$ but not directly on firms in F being Cournot firms. Assume A1* and consider, for example, firms in F behaving in a symmetric conjectural variation mode. Equilibrium for F is characterized now by the symmetric first-

order condition

$$(10) \quad P(Q_M + Q_F) + P'(Q_M + Q_F)Q_F\gamma/k - c = 0$$

where $(\gamma - 1)$ is the conjecture by each firm in F on other firms' responses to its change in output. Thus, $\gamma = 0$ is the competitive case, $\gamma = 1$ is the Cournot case, and $\gamma > 1$ is the case when firms in F expect other firms to expand output in response to an increase in output by F firms. From (10), $1 + dQ_F/dQ_M = P'/[P'k/\gamma + (P' + P''Q_F)]$. Further analysis shows that if $(k - 1)P' - P''Q_F \geq 0$, any $\gamma > 0$ satisfies this condition. However, if $(k - 1)P' - P''Q_F < 0$, which is more plausible, any $0 < \gamma < kP'/[(k - 1)P' + P''Q_F] \equiv z$ satisfies this condition. Note that, by A2, $z > 1$ in the last case. Thus, there is a whole range of conjectures for F, including some range where $\gamma > 1$, where contraction of output by M is not desirable while profitable expansions will raise welfare.

III. Summary and Conclusions

The analysis here seems to suggest, under quite general conditions, that profitable horizontal mergers that start with less than 50 percent of premerger market share are welfare-enhancing. However, a few important caveats are in order.

1) The analysis here ignores income distribution. Expansionary profitable mergers reduce market price and benefit consumers but reduce the profits of the fringe. Contractionary profitable mergers, which are possible when cost-reduction incentives exist, hurt consumers and benefit the fringe. Theorem 1 suggests that expansionary mergers are more likely.

2) To simplify, I did not model the merger process; instead, I implicitly assumed that firms being acquired react passively. Morton Kamien and Israel Zang (1988, 1990) model such processes where firms behave strategically with respect to such activity. They find only limited scope for such mergers and show that, with identical constant marginal

⁹If after the merger there is a sequential game in which M (who deviates from the original equilibrium) moves first, then in a perfect equilibrium of this game M behaves à la Stackelberg.

¹⁰Levin (1988) provides a direct proof that a Cournot firm that becomes a Stackelberg leader will raise its output. However, the analysis there assumes no change in the total number of firms.

cost, such mergers will not take place if mergers of firms with more than 50 percent of market share are prohibited.¹¹

3) Fixed costs that can be shared provide important incentive for mergers. Such costs introduce discontinuity in the cost function and in general may "destroy" existence and uniqueness of the CNE. More important to the analysis is that such fixed costs introduce discontinuity in the function $Q_F(Q_M)$. When Q_M increases beyond Q_M^* , firms in F may reach zero profits with strictly positive levels of output. It is possible then, that expansion in Q_M may cause reduction in total output and an increase in market price. Thus, while Theorem 1 is still valid (if one ignores the existence issue of the original CNE), Theorems 2, 3, and 4 are not valid in general.

4) My analysis assumed that the F remains Cournot. If the fringe changes its behavior after the merger and becomes (say) more cooperative with the M, conclusions may change. This situation will be the subject of a future study.

¹¹Kamien and Zang (1988, 1990) assume that the merged entity remains Cournot. Their results, especially the limited scope for such mergers, depend heavily on this assumption. Also, welfare analysis is absent in their article.

REFERENCES

- Al-Nowaihi, A. and Levine, P. L., "The Stability of the Cournot Oligopoly Model: A Re-assessment," *Journal of Economic Theory*, April 1985, 35, 307-21.
- Bergstrom, Theodore C. and Varian, Hal R., (1985a) "When Are Nash Equilibria Independent of the Distribution of Agents' Characteristics?," *Review of Economic Studies*, October 1985, 52, 715-3.
- _____ and _____ (1985b), "Two Remarks on Cournot Equilibria," *Economics Letters*, 1985, 19, 5-8.
- Farrell, Joseph and Shapiro, Carl, "Horizontal Mergers: An Equilibrium Analysis," *American Economic Review*, March 1990, 80, 107-26.
- Gaudet, Gerard and Salant, Stephen, "Toward a Theory of Horizontal Mergers," in George Norman and Manfredi LaManna, eds., *The New Industrial Economics: Recent Developments in Industrial Organization, Oligopoly and Game Theory*, draft dated July 1989.
- Hahn, Frank H., "The Stability of the Cournot Oligopoly Solution," *Review of Economic Studies*, October 1962, 29, 329-31.
- Kamien, Morton I. and Zang, Israel, "Competitively Cost Advantageous Mergers and Monopolization," working paper, Northwestern University, December 1988.
- _____ and _____, "The Limits of Monopolization Through Acquisition," *Quarterly Journal of Economics*, May 1990, 2, 465-99.
- Levin, Dan, "Cournot Oligopoly and Government Regulation," the second essay in an unpublished Ph.D. Dissertation, Massachusetts Institute of Technology, 1982, 68-101.
- _____, "Stackelberg, Cournot and Collusive Monopoly: Performance and Welfare Comparisons," *Economic Inquiry*, April 1988, 26, 317-30.
- Rosen, B., "Existence and Uniqueness of Equilibrium Points for Concave N -Person Games," *Econometrica*, July 1965, 33, 520-34.
- Ruffin, Roy, "Cournot Oligopoly and Competitive Behavior," *Review of Economic Studies*, October 1971, 38, 493-502.
- Salant, Stephen W., Switzer, Sheldon and Reynolds, Robert J., "Losses From Horizontal Merger: The Effects of an Exogenous Change in Industry Structure on Cournot-Nash Equilibrium," *Quarterly Journal of Economics*, May 1983, 2, 185-99.
- Scherer, F. M., *Industrial Market Structure and Economic Performance*, 2nd ed., Chicago: Rand McNally, 1980.
- Varian, Hal R., *Microeconomic Analysis*, 2nd ed., New York: Norton, 1984.
- White, Lawrence J., "Antitrust and Merger Policy: Review and Critique," *Journal of Economic Perspectives*, Fall 1987, 1, 13-22.

Input Market Price Discrimination and the Choice of Technology

By PATRICK DEGRABA*

Recent concerns over the effects of the Robinson-Patman Act¹ and so-called price-protection policies such as most-favored-customer clauses (MFC's)² on market performance have given economists new reasons to examine the welfare effects of third-degree price discrimination. In order to assess these effects correctly, it is imperative that one understand how price discrimination influences market behavior.

Joan Robinson's (1933) work launched the formal inquiry into the welfare effects of third-degree price discrimination. Building on the intuition presented by Arthur Pigou (1932), she showed that, if a monopolist faces two independent linear demand curves, the use of price discrimination will not affect industry output but will reduce welfare. Richard Schmalensee (1981) extends these results to nonlinear demand curves and shows that an increase in total industry output is a necessary condition for price discrimination to be welfare improving. Hal Varian (1985) broadens these results by deriving upper and lower bounds on the welfare change due to the use of price discrimination. He shows that these results can be applied to markets in which there are nonzero cross price effects.

All of this work examines how the ability of a monopolist to price-discriminate will affect the market outcome when all other characteristics of the market are treated as exogenous. Recently, two lines of research have extended this inquiry beyond the case

of a monopolist in a market with exogenously fixed parameters.

The first line considers the case of oligopoly. The work of Charles Holt and David Scheffman (1985) and Thomas Cooper (1986) has shown that restrictions on price discrimination imposed by the use of MFC's can facilitate collusion between oligopolists attempting to restrict output. This implies that third-degree price discrimination can be welfare-improving.

The second line of research shows that price discrimination by a firm can affect nonprice decisions made by other market participants, thus affecting the market outcome. Michael Katz (1987) presents a model in which a large firm's ability to vertically integrate backward into the production of an input allows it to obtain a lower per-unit price from the supplier of the input than can be obtained by smaller firms without this ability. He shows that third-degree price discrimination reduces welfare unless it prevents inefficient backward integration. DeGraba (1987) shows that the use or nonuse of price discrimination by a national firm can affect nonprice decisions made by local firms that compete with the national firm. In this situation, third-degree price discrimination is welfare-reducing, because it induces local firms to produce a product that is "overly differentiated" from the product of the national firm.

In all of the work cited above, price discrimination is important when sellers set prices in separate markets or charge different prices to different customers in the same market. The following analysis (which can be considered a contribution to the second line of research) suggests that price discrimination can be important even when a seller faces a single market in which all customers are identical. The intuition behind this result is that nonprice decisions made by downstream producers (such as the choice of technology) can be affected by the use or

*Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853. I thank Robert Frank, Robert Smiley, Richard Thaler, and the participants of the JGSM applied microeconomics workshop for their helpful comments.

¹See William Baldwin (1987 pp. 438-40) for a good summary of the debate.

²See John Kwoka and Lawrence White (1989 pp. 196-7).

nonuse of price discrimination by the supplier of an input. Even though the downstream producers will reach a symmetric equilibrium, the equilibrium that they reach under price discrimination will be different from the one they reach under uniform pricing. This means that the pricing strategy of the supplier can determine which symmetric equilibrium producers choose.

I present a simple model to examine how price discrimination in a market for a variable input affects downstream producers' long-run choice of a production technology. In the model, a monopoly supplier of a variable input sells to two downstream producers, who use the input in the production of a final good. These producers must first choose a level of marginal cost (a lower marginal cost is obtained by incurring a higher fixed cost) and then compete as Cournot duopolists who face a linear demand curve in the final goods market. I compare the market equilibrium in which the supplier is allowed to price-discriminate to the equilibrium in which the supplier must charge a uniform price to both downstream firms. This comparison yields two main results.

- 1) If the downstream producers have different constant marginal costs of production, the price-discriminating input supplier will charge the low-cost producer a higher price than he charges the high-cost producer, partially offsetting the cost advantage.
- 2) The downstream producers will choose a technology with a higher marginal cost when the supplier price-discriminates than they will if the supplier charges a uniform price. Further, if the trade-off between marginal cost and fixed costs is a quadratic relationship, welfare in the long run is lower under price discrimination.

I. The Short Run

In the formal model, I present a two-stage three-player game. The players include a monopoly supplier of a variable input and two downstream producers of a final good. The supplier sells the input to the produc-

ers, who use this input along with others to produce a homogeneous output. Units are normalized so that one unit of the input is required to produce one unit of the output.

In the first stage of the game, the monopoly supplier quotes a per-unit price, k_i , for the input to each downstream firm, i . In stage 2, the downstream firms observe these prices and then compete as Cournot duopolists who face a linear market demand curve. Firm i 's per-unit cost of output is $k_i + c_i$, where c_i is an additional marginal cost of production. It may be helpful to think of c_i as the cost of using other variable inputs which may differ across firms due to such reasons as different geographic location or the use of different technologies. I assume that each c_i is set in a competitive market, so it represents the true cost of the resources used.

The strategy for the supplier is an ordered pair $(k_1, k_2) \in \mathbb{R}^{2+}$, where k_1 is the price quoted to firm 1 and k_2 is the price quoted to firm 2. The strategy for each firm, i , is a function, $Q_i: \mathbb{R}^{2+} \rightarrow \mathbb{R}^{1+}$. This function maps every possible observed combination of k_1 and k_2 into a quantity choice.

The payoff to the supplier is $\pi_s = \sum_i k_i q_i$, which is the revenue generated from the sales of the input. (The supplier produces the input at zero marginal cost). Let p be the price of the final good. Producer i 's payoff is $\pi_i = (p - k_i - c_i)q_i$, which is the net revenue from sales in the final goods market.

The equilibrium concept employed is that of subgame perfect Nash equilibrium (Reinhard Selten, 1975). An equilibrium strategy choice is an ordered quadruple $(k_1^*, k_2^*, Q_1^*, Q_2^*)$ such that: 1) no player could improve his payoff by unilaterally deviating and 2) (Q_1^*, Q_2^*) constitute Nash equilibrium choices of q_1 and q_2 , for every possible choice of (k_1, k_2) .

The calculation of the perfect Nash equilibrium proceeds in two steps. The first step is to note that, in stage 2, the producers simply choose the (unique) Cournot output levels, (q_1^*, q_2^*) , for each subgame defined by treating $k_1 + c_1$ and $k_2 + c_2$ as the marginal costs of production.

Once this has been done, the first stage of the game can be viewed as a profit-maximi-

zation problem for the supplier in which he simply chooses (k_1, k_2) to maximize $\pi_s = k_1 Q_1^* + k_2 Q_2^*$. The payoff to each of the producers then is the profit of the subgame defined by the supplier's choice of prices.

When the market demand curve is of the form

$$(1) \quad p = a - b(q_1 + q_2)$$

where $a, b > 0$, the Nash equilibrium strategies for the producers are given by

$$(2a) \quad Q_1^* = (a - 2c_1 - 2k_1 + c_2 + k_2)/3b$$

$$(2b) \quad Q_2^* = (a - 2c_2 - 2k_2 + c_1 + k_1)/3b.$$

The structure outlined thus far can be used to define two games, Γ^d and Γ^u . In Γ^d the supplier is able to price-discriminate, and in Γ^u the supplier is constrained to charge a uniform price. When the supplier is allowed to price-discriminate, a simple maximization calculation reveals that the equilibrium values of k_1 and k_2 are

$$(3a) \quad k_1^{d*} = (a - c_1)/2$$

$$(3b) \quad k_2^{d*} = (a - c_2)/2.$$

Likewise, when the supplier is constrained to charge the same price to both firms, a constrained maximization calculation yields

$$(4) \quad k^{u*} = (2a - c_1 - c_2)/4.$$

OBSERVATION 1: *When the supplier is allowed to price-discriminate, he charges the firm with the lower marginal cost a higher price than he charges the firm with the higher marginal cost. This price differential is less than (the absolute value of) the marginal cost differential.*

This can be seen by setting $c_1 < c_2$ (without loss of generality) in (3a) and (3b). Katz (1987) presents this result for Cournot players facing any demand curve that has a

downward-sloping marginal-revenue curve.³ The reason for this result is clear. The firm with a lower marginal cost has the more inelastic demand for the input, which causes the supplier to charge him a higher price. The result is easily seen in this example, because the demand for the input is a linear function of the prices of the inputs. The additional marginal cost of production, c_i , affects this demand function only through the constant term, and it does so with a negative sign. Thus, a lower c_i causes the demand for the input to be more inelastic, which leads to a higher price.

The general restrictions on downstream competition under which $k_1^{d*} > k^{u*} > k_2^{d*}$ if and only if $c_1 < c_2$, are not known.⁴ It is easy to show that this relationship holds if the demand curves of downstream firms are linear and symmetric. Note that this includes the Cournot example above as well as price-setting games in which the demand in the final goods market is linear (the latter set of games includes the case in which firms are symmetrically placed along a Hotelling line, where all consumers have the same reservation price for the good, as a special case). Note that since k_1^{d*} , k_2^{d*} , and k^{u*} are derived from interior equilibrium points, the relationship $k_1^{d*} > k^{u*} > k_2^{d*}$ if and only if $c_1 < c_2$, must hold for (at least) "small" uniformly continuous perturbations of both the symmetry of the demand curves and the linearity.

³This result implies that the firm that purchases more of the input pays a higher price. This might seem to contradict the more intuitive notion that larger users tend to pay less than small users. The apparent contradiction stems from the fact that quantity discounts are used as a self-selection mechanism when the seller does not know the demand curves of the buyers. In the example, the seller does know the demand curve for each buyer, so quantity discounts are unnecessary.

⁴In fact, the conditions under which a monopolist's optimal discriminatory prices will bracket his optimal uniform price are not completely known. DeGraba (1989) provides some general restrictions on the monopolist's profit functions for which this is true. Similar results for the case of independent markets were developed simultaneously and independently by Babu Nahata et al. (1990).

OBSERVATION 2: *When the supplier price-discriminates, the low-marginal-cost producer produces less, and the high-marginal-cost producer produces more than they would under uniform pricing. In the short run, therefore, welfare is lower under price discrimination.⁵*

PROOF:

Direct calculation shows that, with price discrimination, firm i produces

$$(5a) \quad q_i^d = (a - 2c_i + c_{-i})/6b$$

while with a uniform price, the low-cost firm produces

$$(5b) \quad q_i^u = [a - (7/2)c_i + (5/2)c_{-i}]/6b$$

where $-i$ indicates firm i 's opponent.

It is obvious from inspection that, if $c_i < c_{-i}$, then $(5b) > (5a)$ which implies that the low-cost firm produces less when there is price discrimination. Since total industry output is unaffected by the choice of the supplier's pricing policy (a direct result of linear demand), the high-cost firm must produce more. This means that discriminatory input prices raise the total cost to society of producing the equilibrium quantity of output. This is welfare-reducing.

The lesson to be learned from Observations 1 and 2 is simple. When a price-discriminating supplier of an input sells to producers who have different marginal costs, there is an incentive for the supplier to charge a higher price to the lower-cost firm. As a result, the supplier partially "subsidizes" the cost differential between the two firms. When the supplier sets a uniform price, he does not provide this subsidy. Thus, price discrimination results in a smaller cost differential between the two firms, which causes the lower-cost firm to produce less

and the higher-cost firm to produce more than under uniform pricing. This fact will drive the results of the next section.

II. The Long Run

In this section, I present a model that suggests that the pricing policy of the supplier can affect decisions made by downstream producers other than the short-run choice of quantity or price. Specifically, each producer faces the task of choosing his production technology, which will affect the firm's cost structure and, therefore, his ability to compete against his rivals. I show that when the supplier charges discriminatory prices, the producers choose a technology with a higher marginal cost than they would when the supplier sets a uniform price.

I extend the model of Section I by allowing each downstream producer to choose his level of marginal cost. A lower marginal cost, c_i , can be obtained at the expense of a higher fixed cost, F_i . In order to obtain a closed-form solution, I use the function, $F_i = \alpha c_i^2 - \beta c_i + \gamma$, for $0 \leq c_i \leq \beta/2\alpha$. As with the c_i 's, I assume that the F_i 's represent the true cost of the fixed resources used by the producers.

The following restrictions must now be placed on the parameters:

- (R1) $\alpha, \beta > 0$
- (R2) $(1/9b) - \alpha < 0$
- (R3) $(7/4)(a/9b) - \beta < 0$
- (R4) $a\alpha - (\beta/2) > 0$
- (R5) $\gamma < a^2/36b$.

Inequality (R1) implies that F_i is downward-sloping and convex for $0 \leq c_i \leq \beta/2\alpha$. Inequality (R2) is the statement that the second derivative of profit with respect to c is negative. Expressions (R3) and (R4) are sufficient to guarantee that the first-order conditions yield interior solutions: (R3) states that F_i is steeper than the net revenue function (π_i^u , defined as sales minus variable costs) when the supplier is setting a

⁵In general, there will be two welfare effects, one resulting from reallocating production among different-cost producers and the other resulting from a change in total output. It is difficult to make statements about welfare when these two effects conflict.

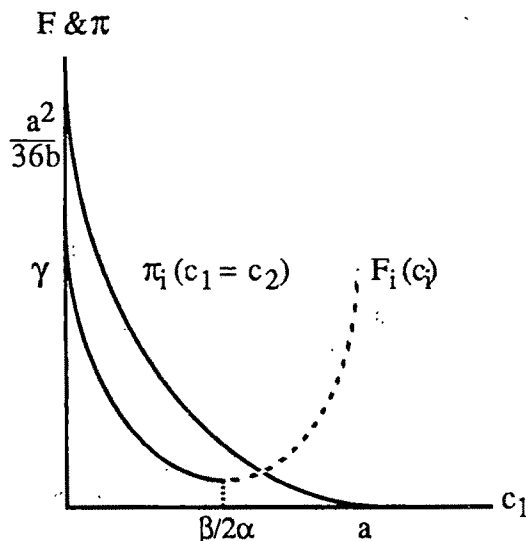


FIGURE 1

uniform price and $c_1 = c_2 = 0$; (R4) is the statement that F_i reaches a minimum at a value of $c_i < a$ (see Fig. 1). Finally, (R5) is a sufficient (but by no means necessary) condition for producers to earn a strictly positive profit. It says that the fixed costs associated with a zero marginal cost are less than the net revenues earned at $c_i = 0$. This is clearly a much stronger condition than necessary, but it is simple and is sufficient for my purposes.

I can now present a three-stage game. The producers choose c_i in stage 1. I assume that this choice is made with the knowledge of whether or not the supplier will employ a uniform price. This assumption can be justified on the grounds that offering or not offering an MFC has been a customary practice in the past. An example of this might be a labor union pledging to negotiate the same wage contract with all manufacturers in a given industry. There may also be laws such as the Robinson-Patman Act, which effectively prevents a supplier from price-discriminating among downstream competitors.

Once the technology decisions have been made, the game proceeds as described in Section I. The supplier offers a price to each firm, and the firms choose quantity

taking their technology choices and input prices as given.

Formally, the strategy of the supplier is a function $K: \mathbb{R}^{2+} \rightarrow \mathbb{R}^{2+}$. K maps every possible ordered pair of (c_1, c_2) into input price choices (k_1, k_2) . The strategy of producer i has the form (c_i, Q_i) , where the number $c_i \in [0, \beta/2\alpha]$ is his marginal cost of production, and the function $Q_i: \mathbb{R}^{2+} \rightarrow \mathbb{R}^{1+}$, maps $(c_1 + k_1, c_2 + k_2)$ into a quantity choice, q_i . The supplier's payoff is $\pi_s = \sum_i k_i q_i$, and the payoff to producer i is

$$(6) \quad \pi_i = [a - b(q_1^* + q_2^*) - c_i - k_i]q_i^* - [ac_i^2 - \beta c_i + \gamma].$$

The equilibrium concept is that of perfect Nash equilibrium.

PROPOSITION 1: *The technology chosen by producers under price discrimination has a higher marginal cost than the technology chosen under uniform pricing. Output in the final goods market is therefore lower under price discrimination than under uniform pricing.*

PROOF:

As with all multistage games, the appropriate procedure is to solve the game backwards. The stage 2 and 3 solutions have already been calculated in the previous section. It is only necessary to solve the first stage of this game.

To calculate the choices of c_1 and c_2 when the supplier is allowed to price-discriminate, equations (3a), (3b), and (5a) must be substituted into (6). Solving the first-order conditions simultaneously yields

$$(7a) \quad c_1^d = c_2^d = (a - 9b\beta)/(1 - 18b\alpha)$$

along with the other equilibrium values:

$$(7b) \quad k_1^d = k_2^d = -9b(2a\alpha - \beta)/2(1 - 18b\alpha)$$

$$(7c) \quad q_1^d = q_2^d = -3(2a\alpha - \beta)/2(1 - 18b\alpha).$$

When the supplier is constrained to charge a uniform price, equations (4) and (5b) are substituted into (6). Solving these first-order conditions yields:

$$(8a) \quad c_1^u = c_2^u = \frac{(7/4)a - 9b\beta}{(7/4) - 18b\alpha}$$

$$(8b) \quad k_1^u = k_2^u = \frac{-9\gamma(2a\alpha - \beta)}{2[(7/4) - 18b\alpha]}$$

$$(8c) \quad q_1^u = q_2^u = \frac{-\gamma(2a\alpha - \beta)}{2[(7/4) - 18b\alpha]}.$$

From the restrictions placed on the parameters, we know that $(7/4)a - 9b\beta < 0$, $(7/4) - 18b\alpha < 0$, and $2a\alpha - \beta > 0$. These inequalities guarantee that (7a)–(7c) and (8a)–(8c) are positive. It is then simple to show that $c_1^u < c_1^d$ if and only if $a\alpha - \beta/2 > 0$, and the fact that $q_1^u > q_1^d$ is obvious from inspection.

This model suggests that, when the supplier price-discriminates, the producers will choose a technology with a higher marginal cost than they will when the supplier charges a uniform price. The reason is simple. It was seen earlier (Observation 1) that if there is a difference in marginal costs, the discriminating supplier will charge the producer with a low marginal cost a higher input price, partially offsetting the cost differential.

In this case, the price-discriminating supplier reduces the incentives to reduce marginal cost unilaterally, because the advantage gained through achieving a lower marginal cost is partially offset by the differential in the supplier's input prices. When the supplier charges a uniform price, each firm receives the full benefit of a cost reduction (since none of the advantage is dissipated through price discrimination). Thus, unilateral cost reductions that are marginally profitable under uniform pricing are unprofitable under price discrimination. This results in a higher equilibrium marginal cost under price discrimination. The higher marginal cost then translates into a lower level of output in the final goods market.

While the results are shown for the simple linear case above, the intuition can be generalized as follows. Consider the game in which each competitor chooses a level of c_i . If the best-response functions of this game are either downward-sloping or upward-sloping with slope less than 1 and the monopolist finds it optimal to charge the low-cost firm more than he charges the high-cost firm, then price discrimination results in a higher choice of c and lower output than does the use of uniform pricing.⁶

It should be noted that, unlike the results in the short run, the supplier's profit in the long run is lower under price discrimination than it would be under uniform pricing. The reason is that discriminatory prices induce the firms to choose a higher marginal cost, which causes the supplier to charge a lower price. Since less output is produced, the supplier sells less of the input at the lower price, resulting in a lower profit.

PROPOSITION 2: *Welfare (as measured by the sum of consumer and producer surplus) is lower under price discrimination than under uniform pricing.*

PROOF:

I prove this proposition using two lemmas.

LEMMA 1: *Each producer's average cost (in terms of real resources used) of producing q^u under uniform pricing (AC^u) is less than his average cost of producing q^d under price discrimination (AC^d)*

PROOF:

See the Appendix.

LEMMA 2: *In equilibrium, under uniform pricing, the price in the final goods market is*

⁶See DeGraba (1988) for the proof. If best-response functions have slopes between -1 and 1 , then the equilibrium will be stable, as in Avinash Dixit (1986). The result can be extended further to games with best-response functions with slope less than -1 , but the interpretation of such an equilibrium is troublesome.

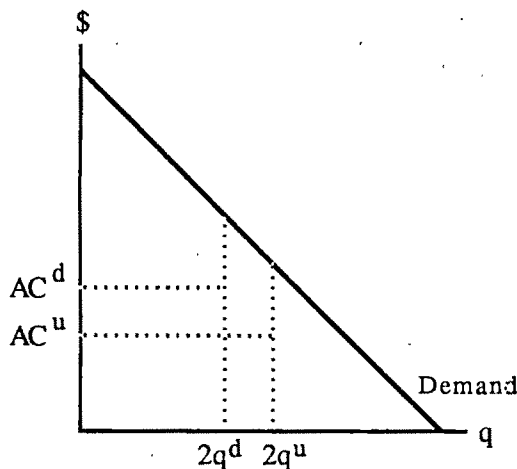


FIGURE 2

greater than the producer's average cost of producing q^u .

PROOF:

This follows immediately from the fact that each firm earns a positive profit, which is guaranteed by (R5).

Lemmas 1 and 2 can be used to generate the graph in Figure 2 for the final goods market equilibria. This graph clearly shows that welfare under price discrimination is lower than welfare under uniform pricing.

The proof suggests that the use of price discrimination creates two effects which decrease social welfare. The first is the fact that output is lower under price discrimination. This is not surprising. Decreasing output virtually always decreases welfare in Cournot models.

The second is that price discrimination changes the marginal-cost/fixed-cost "mix." Under price discrimination, producers choose a level of c that is too high (and correspondingly a level of F that is too low) from a welfare point of view. The use of uniform pricing causes producers to choose a lower c and a higher F , which in this situation is more efficient.

To illustrate, consider a producer who must choose c to minimize the following

cost function:

$$(9) \quad TC = \alpha c_1^2 - \beta c_1 + \gamma + c_1 q_1 + k_1 q_1$$

where all parameters are positive, k_1 is *not* a function of c_1 , and q_1 is fixed. For simplicity, let the input for which k_1 is the price, have a zero cost of production. The first-order conditions tell us that the cost-minimizing c_1 is given by

$$(10) \quad c_1^* = (\beta - q_1)/2\alpha.$$

Notice that this choice of c_1^* also minimizes society's cost of producing q_1 .

Now let k_1 be a (decreasing) function of c_1 . This will alter the first-order conditions from (9) so that the firm's optimal choice of c_1 no longer minimizes society's cost of production. Substituting (3a) into (9) and (4) and (9) and optimizing, one obtains

$$(11a) \quad c_1^d = [\beta - (1/2)q_1]/2\alpha$$

$$(11b) \quad c_1^u = [\beta - (3/4)q_1]/2\alpha$$

respectively. Notice that c_1^d is farther from c_1^* than is c_1^u . This is because the use of discriminatory prices "augments" the effect of c_1 on k_1 vis-à-vis uniform pricing, which increases the distortion caused by this functional relationship. Thus, price discrimination causes firm 1 to choose c farther from c_1^* , thereby increasing the cost of production and reducing welfare.

The message from this analysis is clear. When a supplier finds it in his interest to charge low-cost producers a high price and high-cost producers a low price, price discrimination creates incentives that discourage the reduction of marginal cost at the expense of fixed cost. This results in an industry output that is lower than would occur under uniform pricing. This decrease in output, along with the use of a less efficient marginal cost/fixed-cost mix, produces welfare that is lower than would occur under uniform pricing.

APPENDIX

PROOF OF LEMMA 1: Consider a firm that has chosen the technology with marginal cost c^d and a firm which has chosen the technology with marginal cost c^u . Suppose they each produce q^u units. A fair amount of algebra is required to show that the total cost for the firm with technology c^u is less than the total cost for the firm with technology c^d . Subtracting the total cost of production of the firm with marginal cost c^u from the total cost of the firm with marginal cost c^d yields

$$(A1) \quad \Delta TC = \alpha(c^d)^2 - \beta c^d + c^d q^u - [\alpha(c^u)^2 - \beta c^u + c^u q^u].$$

Plugging (7a), (7c), (8a), and (8c) into (A1) and simplifying yields

$$(A2) \quad \Delta TC = [(1/\varepsilon)(9b\alpha + 1)(2a\alpha - \beta)].$$

This is clearly positive. Since both firms are producing the same output level, the firm with marginal cost c^d also has a higher average cost of production.

Now let the firm with marginal cost c^d reduce its output from q^u to q^d . Since $(F + cq)/q$ is a decreasing function of q for $F, c > 0$, decreasing output increases average cost. Therefore, a firm producing q^d with technology c^d has a higher cost than a firm producing q^u with technology c^u .

REFERENCES

- Baldwin, William, L., *Market Power, Competition, and Anti Trust Policy*, Homewood, IL: Irwin, 1987.
- Cooper, Thomas, "Most-Favored Customer Pricing and Tacit Collusion," *Rand Journal of Economics*, Autumn 1986, 17, 377-88.
- DeGraba, Patrick, "The Effects of Price Restrictions on Competition Between National and Local Firms," *Rand Journal of Economics*, Autumn 1987, 18, 333-47.
- _____, "Input Market Price Discrimination and the Choice of Technology," Johnson Graduate School of Management Working Paper, 1988.
- _____, "The Relationship Between Optimal Third Degree Discriminatory Prices and the Optimal Uniform Price," Johnson Graduate School of Management Working Paper, 1989.
- Dixit, Avinash, "Comparative Statics for Oligopoly," *International Economic Review*, February 1986, 27, 107-22.
- Holt, Charles and Scheffman, David, "The Effects of Advance Notice and Best Price Policies: Theory with Applications to Ethyl," Federal Trade Commission, Bureau of Economics Working Paper No. 106, 1985.
- Katz, Michael, "The Welfare Effects of Third-Degree Price Discrimination in Intermediate Good Markets," *American Economic Review*, March 1987, 77, 154-67.
- Kwoka, John and White, Lawrence, *The Antitrust Revolution*, Glenview, IL: Scott, Foresman, 1989.
- Nahata, Babu, Ostaszewski, Krzysztof and Sahoo, P. K., "Direction of Price Changes in Third-Degree Price Discrimination and Some Welfare Implications," *American Economic Review*, 1990, 80, 1254-8.
- Pigou, Arthur, *The Economics of Welfare*, 4th Ed., London: Macmillan, 1932.
- Robinson, Joan, *Economics of Imperfect Competition*, London: Macmillan, 1933.
- Schmalensee, Richard, "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, 71, 242-7.
- Selten, Reinhard, "Re-examination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 1975, 4, 25-55.
- Varian, Hal, "Price Discrimination and Social Welfare," *American Economic Review*, September 1985, 75, 870-5.

Direction of Price Changes in Third-Degree Price Discrimination

By BABU NAHATA, KRZYSZTOF OSTASZEWSKI, AND P. K. SAHOO*

There is a long tradition in economic literature of comparing welfare under discriminatory pricing with nondiscriminatory pricing employing the usual Marshallian measure. For example, Richard Schmalensee (1981), by assuming independent demand and constant marginal cost, clearly shows that if welfare were to increase under discrimination, then the total output with discrimination must be higher than the total output without discrimination. Hal Varian (1985) analyzes the welfare question in a more general setting. He develops bounds that serve as a sufficient condition for welfare to increase. In addition, these bounds provide important insights in predicting the welfare change under different demand and cost functions. Marius Schwartz (1990) proves that, for any cost function, if discrimination decreases output, it will decrease welfare, as well.

There are two reasons why the welfare question has been so difficult to answer. 1) Based on his analysis Schmalensee (1981 p. 245) concludes that "If all demand functions are strictly concave or convex and if the p_i 's [prices in each submarket after discrimination] are not nearly equal, there is apparently no simple, general way to tell if monopolistic discrimination will raise or lower total output."¹ Given this absence of

a simple test, it is not possible to verify whether the necessary condition (i.e., increase in output under discrimination) is satisfied or not. 2) Even if one were able to develop a test that could predict when the total output would increase under discrimination, it may not be very useful in answering the basic question of welfare change, because an increase in total output is only a necessary condition and hence does not guarantee an increase in welfare. Thus, to answer the welfare question, a new approach is needed for theoretical analysis of third-degree price discrimination.

Instead of concentrating on the output effects of discrimination, this paper focuses on the price effects of discrimination. A basic result that has remained unquestioned in the literature is that when there are two classes of buyers, discrimination raises price for one class and lowers it for the other. However, in an interesting paper focusing on price discrimination in intermediate good markets, Michael Katz (1987 p. 156) concludes that "Under reasonable conditions, intermediate good price discrimination leads to higher input price being charged to *all* buyers, a result that never would arise in a corresponding model of a final good market." According to Katz this surprising result is a striking example of the difference between intermediate and final good markets. We not only show that discrimination in the final good market can also raise prices for all buyers (the result denied by Katz),

*The first author is in the Department of Economics and Finance, and the last two authors in the Department of Mathematics, University of Louisville, Louisville, KY 40292. Comments of Stephen Layson and Marius Schwartz were very helpful in revising the paper. The geometrical explanation for our results using Figures 1 and 2 is due to Stephen Layson, for which we are thankful. We also thank other anonymous referees. We acknowledge the research assistance of Zhao-Bi Ha. Our research was partially supported by the University of Louisville. An earlier version of this paper was presented at the Fourth Congress of the European Economic Association in Augsburg, West Germany, in September 1989.

¹Joan Robinson (1933 pp. 192-5) proposed a test based on the adjusted concavities of the submarkets' demand functions at the nondiscriminating monopoly price to determine whether total output rises or falls after discrimination. However, Melvin Greenhut and Hiroshi Ohta (1976), with the help of an example, have clearly demonstrated that Robinson's proposed test is not valid under all conditions and, hence, seems to have little real value.

but further show that it can also *lower* the prices for all buyers. It is important to emphasize that when prices move in the same direction in both markets, the welfare effect of discrimination may be quite large compared to the more typical situation when price rises in one market and falls in the other. When both prices move in the same direction, the welfare effects are predictable. When both prices go down, consumer's surplus increases because of the price movement, while profit increases due to discrimination; thus, welfare must go up. When prices go up, the total output is reduced, which causes welfare to decrease.

I. Main Results

We first give two examples to demonstrate that price discrimination can either increase prices in both markets or decrease prices in both markets. We do not restrict the demand functions to be strictly concave or convex. We only require that they are continuous and twice differentiable with negative slope throughout their range. We also do not restrict the profit functions to be strictly concave or restrict marginal revenue curves to be declining continuously (e.g., Schmalensee, 1981 p. 243). John Formby, Stephen Layson, and James Smith (1982) clearly demonstrate that the assumption of continuously declining marginal revenue may be quite restrictive, and "...demand conditions leading to upward sloping marginal revenue may indeed be pervasive" (Formby et al., 1982 p. 306). Based on their examples they conclude that, "...very simple analytical demand curves may have non-trivial upward sloping marginal revenue curves and that multiple profit equilibria for firms cannot be easily dismissed" (p. 309). We use polynomial demand functions to illustrate these results for the following reasons. First, a polynomial demand function relaxes the assumption of strict concavity or convexity, and at the same time, it also relaxes the assumptions of declining marginal revenue and concavity of profit functions. Second, any sufficiently smooth demand function can be approximated with a polynomial using Taylor's expansion with

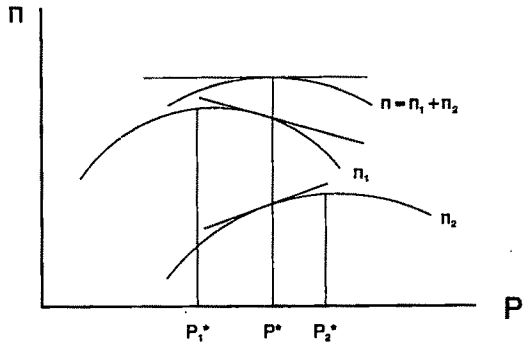


FIGURE 1. OPTIMAL PRICES WITH CONCAVE PROFIT FUNCTIONS

any desired degree of accuracy. Thus, a polynomial demand function, being more general, can provide an analytical framework for identifying other types of demand functions where similar results may hold. For the sake of illustration, we consider only two markets and constant marginal cost. Both markets are served with and without discrimination.

Figures 1 and 2 provide geometrical explanation for our results. If $\pi_1(p)$ and $\pi_2(p)$ are profit functions for submarkets 1 and 2, respectively, and $\pi(p) = \pi_1(p) + \pi_2(p)$ is the uniform-price profit function, then at the single uniform profit-maximizing price, p^* , $\pi'(p) = \pi'_1(p) + \pi'_2(p) = 0$. Note in Figure 1 that, at p^* , $\pi'_1(p) < 0$ and $\pi'_2(p) > 0$. Therefore, if π_1 and π_2 both are concave, $p_1^* < p^* < p_2^*$, the usual case. If, however, π_2 has two local maxima, then it is possible that $p_1^* < p^*$ and $p_2^* < p^*$ or $p_1^* > p^*$ and $p_2^* > p^*$. The former case is illustrated in Figure 2 (for expository purposes, the shapes of the graphs of three profit functions are more exaggerated for Example 1).

Example 1: Discrimination lowers prices in both markets.

Market 1 demand:

$$Q_1 = -0.25p_1^3 + 2.0001p_1^2 - 5.5p_1 + 10$$

Market 2 demand:

$$Q_2 = -0.2561p_2^3 + 2.7p_2^2 - 9.5p_2 + 12$$

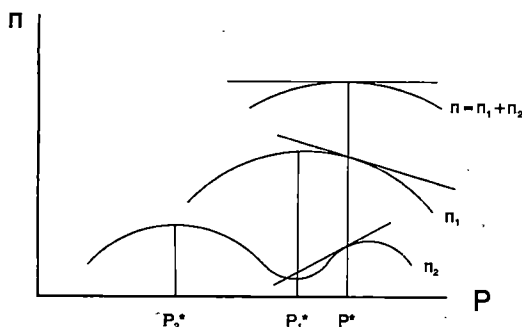


FIGURE 2. OPTIMAL PRICES WITH UNRESTRICTED PROFIT FUNCTIONS

Combined demand:

$$Q = -0.5061p^3 + 4.7001p^2 - 15p + 22$$

Marginal cost:

$$c = 0.1$$

Profit $\pi(p)$ without discrimination:

$$\begin{aligned}\pi(p) &= -0.5061p^4 + 4.7001p^3 - 15p^2 + 22p \\ &\quad - 0.1\{-0.5061p^3 \\ &\quad + 4.7001p^2 - 15p + 22\} \\ &= -0.5061p^4 + 4.75071p^3 \\ &\quad - 15.47001p^2 + 23.5p - 2.2\end{aligned}$$

Profit maximization yields²

$$\begin{aligned}p^* &= 3.859496, \quad Q^* = 5.0232297, \\ \pi^* &= 18.884816, \quad Q_1 = 4.1931976, \\ Q_2 &= 0.8300321.\end{aligned}$$

Profit $\hat{\pi}(p_1, p_2)$ with discrimination:

$$\begin{aligned}\hat{\pi}(p_1, p_2) &= -0.25p_1^4 + 2.0001p_1^3 - 5.5p_1^2 + 10p_1 \\ &\quad - 0.2561p_2^4 + 2.7p_2^3 - 9.5p_2^2 + 12p_2 \\ &\quad - 0.1\{-0.25p_1^3 + 2.0001p_1^2 - 5.5p_1 \\ &\quad + 10 - 0.2561p_2^3 + 2.7p_2^2 - 9.5p_2 + 12\}\end{aligned}$$

²All these values represent global optima. Detailed numerical computations for the two examples are available from the authors upon request.

Profit maximization yields

$$\begin{aligned}p_1^* &= 3.809910 & p_2^* &= 1.097442 \\ Q_1^* &= 4.2521695 & Q_2^* &= 4.4876276 \\ \pi_1^* &= 15.775169 & \pi_2^* &= 4.476147 \\ p^* &> p_1^* > p_2^* & \text{ and } \pi^* < \hat{\pi}^* = \pi_1^* + \pi_2^*.\end{aligned}$$

Note that the output in market 2 increases substantially due to discrimination.

Example 2: Discrimination raises prices in both markets.

Market 1 demand:

$$Q_1 = -0.25p_1^3 + 2.0001p_1^2 - 5.5p_1 + 6$$

Market 2 demand:

$$Q_2 = -0.25p_2^3 + 2.655p_2^2 - 9.5p_2 + 12.5$$

Combined demand:

$$Q = -0.5p^3 + 4.6551p^2 - 15p + 18.5$$

Marginal cost:

$$c = 0.1$$

Profit $\pi(p)$ without discrimination:

$$\begin{aligned}\pi(p) &= -0.5p^4 + 4.6551p^3 - 15p^2 + 18.5p \\ &\quad - 0.1\{-0.5p^3 + 4.6551p^2 - 15p + 18.5\}\end{aligned}$$

Profit maximization yields

$$\begin{aligned}p^* &= 1.158687, \quad Q^* = 6.5916246, \\ \pi^* &= 6.978467, \quad Q_1 = 1.9235665, \\ Q_2 &= 4.6680581.\end{aligned}$$

Profit $\hat{\pi}(p_1, p_2)$ with discrimination:

$$\begin{aligned}\hat{\pi}(p_1, p_2) &= -0.25p_1^4 + 2.0001p_1^3 - 5.5p_1^2 + 6p_1 \\ &\quad - 0.25p_2^4 + 2.655p_2^3 - 9.5p_2^2 + 12.5p_2 \\ &\quad - 0.1\{-0.25p_1^3 + 2.0001p_1^2 - 5.5p_1 \\ &\quad + 6 - 0.25p_2^3 + 2.655p_2^2 \\ &\quad - 9.5p_2 + 12.5\}\end{aligned}$$

Profit maximization yields

$$p_1^* = 3.013776 \quad p_2^* = 1.170637$$

$$Q_1^* = 0.7474162 \quad Q_2^* = 4.616279$$

$$\pi_1^* = 2.177805 \quad \pi_2^* = 4.942359$$

$$p^* < p_2^* < p_1^* \quad \text{and} \quad \pi^* < \hat{\pi}^* = \pi_1^* + \pi_2^*.$$

Note that the output in market 1 decreases significantly due to discrimination.

We now provide sufficient conditions under which the single nondiscriminatory price will always be lower than the maximum of the prices in the submarkets but higher than the minimum of the prices in the submarkets. These results are stated as a theorem and its corollaries.

Consider a monopolist selling a product in a single market. Let $Q = Q(p)$ be the demand function, where p is the price. Let π be the profit function, and let $C(Q)$ be the total cost function. Then,

$$\pi(p) = pQ(p) - C(Q).$$

Now suppose the monopolist segments the market into n distinguishable submarkets, where $n \geq 2$ is a positive integer. Let $Q_i(p)$ be the demand function in i th market, $i = 1, 2, \dots, n$. If the monopolist discriminates, the total profit is

$$\begin{aligned} \hat{\pi}(p_1, p_2, \dots, p_n) &= \sum_{i=1}^n \pi_i(p_i) \\ &= \sum_{i=1}^n p_i Q_i - C\left(\sum_{i=1}^n Q_i\right) \end{aligned}$$

where π_i is the profit in the i th submarket. Let $(p_1^*, p_2^*, \dots, p_n^*)$ be the price vector maximizing the total profit $\hat{\pi}(p_1, p_2, \dots, p_n)$.

If no discrimination occurs, the profit equals $\pi(p) = \hat{\pi}(p, p, \dots, p)$.

THEOREM 1: *If for each $i = 1, 2, \dots, n$, $\pi_i(p_i)$ is, for $p_i > 0$, a continuous function with a global maximum at p_i^* such that, for $p_i < p_i^*$, $\pi_i(p_i)$ is strictly increasing and,*

for $p_i > p_i^$, $\pi_i(p_i)$ is strictly decreasing (so that the profit function is single-peaked), then $(p_1^*, p_2^*, \dots, p_n^*)$ maximizes $\hat{\pi}(p_1, p_2, \dots, p_n)$. If p^* maximizes $\pi(p) = \hat{\pi}(p, p, \dots, p)$, then*

$$\begin{aligned} \min\{p_1^*, p_2^*, \dots, p_n^*\} \\ \leq p^* \leq \max\{p_1^*, p_2^*, \dots, p_n^*\}. \end{aligned}$$

PROOF:

For $p < \min(p_1^*, p_2^*, \dots, p_n^*)$ consider $\hat{\pi}(p, p, \dots, p) = \pi(p) = \sum_{i=1}^n \pi_i(p)$. For $p_i < p_i^*$, each $\pi_i(p_i)$ is an increasing function, so that for $p < \min(p_1^*, p_2^*, \dots, p_n^*)$, $\pi(p)$ is increasing. Similarly, for $p > \max(p_1^*, p_2^*, \dots, p_n^*)$, $\pi(p)$ is decreasing. Also, $\pi(p)$ is a continuous function on the compact interval $[\min(p_1^*, p_2^*, \dots, p_n^*), \max(p_1^*, p_2^*, \dots, p_n^*)]$; thus, it obtains a maximum there. The maximum is global since the values of $\pi(p)$ for p not in $[\min(p_1^*, p_2^*, \dots, p_n^*), \max(p_1^*, p_2^*, \dots, p_n^*)]$ are smaller than for p in the interval. Since $p^* \in [\min(p_1^*, p_2^*, \dots, p_n^*), \max(p_1^*, p_2^*, \dots, p_n^*)]$, this proves the theorem. Note that we make no assumption about the demand and cost functions, but only about the profit function.

Corollaries 1 and 2 are immediate.

COROLLARY 1: *If each profit function π_i is concave, then the conclusion of Theorem 1 holds.*

COROLLARY 2: *If each Q_i , for $i = 1, 2, \dots, n$, is concave (this includes the case of linear demand functions) and c is constant, then the conclusion of Theorem 1 holds (for $p_i > c$).*

COROLLARY 3: *If the demand functions for each market are of the constant-elasticity type and marginal cost is constant, then the conclusion of Theorem 1 holds.*

PROOF:

Consider the i th market for $i = 1, 2, \dots, n$. We have $Q_i(p_i) = p_i^{-\eta_i}$, where $\eta_i > 1$ is a

constant. Then

$$\pi_i(p_i) = (p_i - c)p_i^{-\eta_i}.$$

If $c = 0$, this function does not have a maximum, and in fact $\lim_{p_i \rightarrow 0^+} = +\infty$, so this case is insignificant. If $c > 0$, the function has a unique maximum at $p_i^* = c[\eta_i/(\eta_i - 1)]$, is increasing for $p_i < p_i^*$ and decreasing for $p_i > p_i^*$. Thus, the assumptions of Theorem 1 are met.

Note that the profit functions π_i in the constant-elasticity case are not concave; their concavities change at

$$p_i = c \frac{\eta_i + 1}{\eta_i - 1}.$$

II. Concluding Remarks

In a recent paper, as a corollary to their proposition 2, Jerry Hausman and Jeffrey MacKie-Mason (1988) conclude that, "...if marginal cost is constant, then with more than one market served under uniform pricing, at least one discriminatory price must be higher than the uniform price, so that a Pareto improvement is not possible" (p. 256 and fn. 9, p. 257). Our paper shows that this conclusion in general is not true. If demand functions are not restricted to a particular class, the examples presented above clearly demonstrate that third-degree price discrimination may either lower or raise price in all submarkets. Also, the traditional result that discrimination increases price in some markets and lowers it in the other markets can be obtained.

We emphasize that discrimination results in a Pareto improvement *only* when both discriminatory prices are lower than the uniform price. Hausman and MacKie-Mason (1988) show that, when both markets are served under uniform price, scale economies are necessary for Pareto improvement. They rule out the possibility of Pareto improvement in the absence of scale

economies if both markets are to be served with a uniform price. We show that this need not be the case. Pareto improvement can occur even when marginal cost is constant. Thus, Hausman and MacKie-Mason's conclusion that discrimination can result in a Pareto improvement could be generalized to include the constant marginal cost as well.

Although in this paper we have shown that both prices can move in the same direction as a result of discrimination, the mathematical and economic conditions under which it is true remain unexplored.

REFERENCES

- Formby, John P., Layson, Stephen, and Smith, W. James, "The Law of Demand, Positive Sloping Marginal Revenue, and Multiple Profit Equilibria," *Economic Inquiry*, April 1982, 20, 303-11.
- Greenhut, Melvin L., and Ohta, Hiroshi, "Joan Robinson's Criterion for Deciding Whether Market Discrimination Reduces Output," *Economic Journal*, March 1976, 86, 96-7.
- Hausman, Jerry A., and MacKie-Mason, Jeffrey K., "Price Discrimination and Patent Policy," *Rand Journal of Economics*, Summer 1988, 19, 253-65.
- Katz, Michael L., "The Welfare Effects of Third-Degree Price Discrimination in Intermediate Good Markets," *American Economic Review*, March 1987, 77, 154-67.
- Robinson, Joan, *The Economics of Imperfect Competition*, London: Macmillan, 1933.
- Schmalensee, Richard, "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, 71, 242-7.
- Schwartz, Marius, "Third-Degree Price Discrimination and Output: Generalizing a Welfare Result," *American Economic Review*, December 1990, 80, 1259-62.
- Varian, Hal R., "Price Discrimination and Social Welfare," *American Economic Review*, September 1985, 75, 870-5.

Third-Degree Price Discrimination and Output: Generalizing a Welfare Result

By MARIUS SCHWARTZ*

One of the best-known conjectures in the economics of price discrimination is that a move by a monopolist from uniform pricing to third-degree price discrimination—charging different prices in different exogenously identifiable markets—reduces the sum of consumer surplus and profit (hereinafter “welfare”) if total output decreases. This conjecture can be found, at least implicitly, as far back as A. C. Pigou (1920). It is of some interest, since it suggests a welfare test that only requires knowledge of observable magnitudes. Richard Schmalensee (1981) proves the conjecture assuming that the monopolist can perfectly separate markets and that marginal cost is constant. Hal Varian (1985) extends the result by allowing imperfect arbitrage, so that demand in any market can depend on prices in other markets, and by allowing marginal cost to be constant or increasing. (Schmalensee and Varian establish additional useful results on the welfare effects of third-degree price discrimination.) Using a revealed-preference argument, this note generalizes the result to the case in which marginal cost is decreasing, a serious possibility in the context of monopoly.

In order to motivate the revealed-preference approach, it is helpful to review the intuition for the result when marginal cost is constant or increasing and show why that intuition can break down when marginal cost is decreasing. Suppose that the monopoly output under uniform pricing is q^u and that moving to discrimination yields a total output q^d below q^u . Welfare under

discrimination will be no higher than if the same output q^d is allocated through uniform pricing: uniform pricing allocates a given total output optimally (it leaves no unexploited gains from reshuffling output between markets), while discriminatory pricing in general will induce misallocations by distorting consumers’ choices. Also, welfare achieved if q^d is allocated through uniform pricing will be lower than if the higher output q^u is allocated through uniform pricing. This follows because q^u is the monopolist’s choice under uniform pricing, so the demand curve lies above the marginal cost curve at q^u . If marginal cost is nondecreasing, demand will lie above marginal cost also at lower outputs; hence, reducing output below q^u will reduce welfare.

If marginal cost is decreasing, this type of argument is inconclusive. At some outputs below q^u the demand curve might now lie below the marginal cost curve, as illustrated in Figure 1. Thus, welfare under uniform pricing, $W^u(q)$, can increase over some range as output falls below q^u . I therefore proceed along a different tack, using a revealed-preference argument that relies on q^u being a profit-maximizing output under uniform pricing.

Consider a monopolist selling to n exogenously identifiable markets. Let p_i and q_i respectively denote the price and output sold in market i , $i = 1, \dots, n$. The monopolist’s total cost function is $C(\sum q_i)$; that is, total cost depends only on total output and not on its distribution among markets. The markets can be viewed, for example, as different types of customers (e.g., students, senior citizens), different times of purchase (e.g., lunch vs. dinner), or different locations to which the monopolist ships its output. (In the last case, cost can be independent of the output’s distribution among markets if, for example, markets are

*Georgetown University and Antitrust Division, U.S. Department of Justice. The views expressed here do not necessarily represent those of the U.S. Department of Justice. For helpful discussions and comments, I thank Tim Brennan, Maxim Engers, Martin Richardson, Marilyn Simon, Bert Smiley, and Jean Tirole.

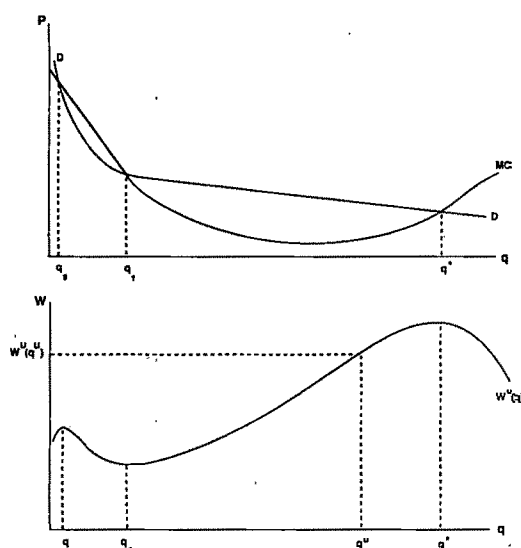


FIGURE 1. WELFARE AND OUTPUT UNDER UNIFORM PRICING:
EXAMPLE IN WHICH WELFARE FUNCTION IS NOT SINGLE-PEAKED

equidistant to the monopolist's plant and transport cost is constant.) Following Varian (1985), I allow imperfect arbitrage among markets (with perfect arbitrage, of course, price discrimination would be impossible). That is, if price differentials are sufficiently high, then goods or customers might move between locations, nonstudents might obtain fake student ID's, and dinner patrons might switch to lunch.

It is not necessary to get into details of the arbitrage technology. One simply thinks of the n markets as representing different goods to consumers and allows each individual's indirect utility function to depend on the prices of all n goods. In order to use the classical welfare measure of total consumer surplus plus profit, each individual's indirect utility function is assumed to be quasi-linear in the vector of n prices and in all other goods, which are treated as a composite commodity y and used as the numeraire. Under the quasi-linear preferences, one can also aggregate across consumers and think of the indirect utility function of a representative individual whose endowment in the numeraire is y_0 : $f(p_1, \dots, p_n, y_0) =$

$v(p_1, \dots, p_n) + y_0$. The function v embodies whatever substitutability exists among the n goods or, equivalently, whatever arbitrage is possible among the n markets. (For discussions of consumer surplus, aggregation, quasi-linear utility, and the composite commodity theorem see Angus Deaton and John Muellbauer [1980] or Varian [1984].)

If the monopolist's n goods are sold under uniform pricing ($p_i = p$ for all i), then one can simplify further and think also of these n goods as a composite commodity whose price is p , and write the indirect utility function as

$$F(p, y_0) = V(p) + y_0.$$

Note that $V(p)$ gives consumer surplus from purchasing the monopolist's composite good at price p (if one normalizes V by setting $V(p) \rightarrow 0$ as $p \rightarrow \infty$). $V(p)$ is always strictly decreasing and weakly convex. Since F is linear in y_0 , the negative of the derivative of V , where it exists, gives the demand function for the composite good: $q = D(p) = -V'(p)$. The only substantive assumption is that $V(p)$ is strictly convex, that is, that the demand for the monopolist's composite good is a strictly decreasing function of price.

Let $W^u(q)$ denote welfare when the monopolist maximizes profit subject to being constrained to charge uniform prices and to sell a given total quantity q :

$$(1) \quad W^u(q) = V(h(q)) + \Pi(q)$$

where $h(q)$ is the inverse demand function and $\Pi(q) = h(q)q - C(q)$ is profit and where, for simplicity, we omit from welfare the endowment term y_0 , which is constant. Observe that V is strictly increasing in q , since it is strictly decreasing in p and since the inverse demand function is strictly decreasing. That is, given a downward-sloping demand curve, consumer surplus is higher if a higher output is sold. Whether pricing is uniform or not, welfare (again ignoring y_0) can also be expressed as utility minus cost:

$$(2) \quad W(q_1, \dots, q_n) = U(q_1, \dots, q_n) - C(\sum q_i).$$

It is now possible to establish the welfare result.

PROPOSITION 1: *Suppose that p^u and $q^u = D(p^u)$ are a profit-maximizing price and output pair when the monopolist is constrained to charge uniform prices. Consider any discriminatory price vector $\mathbf{p}^d = (p_1, \dots, p_n)$, $p_i \neq p_j$ for at least some $i \neq j$, which yields an associated output vector $\mathbf{q}^d = (q_1, \dots, q_n)$, and denote the total output by $q^d = \sum q_i$. If total output is lower under discrimination, then welfare also is lower. That is, if $q^d < q^u$, then $W(\mathbf{q}^d) < W^u(q^u)$.*

PROOF:

I show that $W(\mathbf{q}^d) \leq W^u(q^u) < W^u(q^u)$. Consider the first inequality. For any total output q , let $W^*(q)$ denote the solution to the planner's problem: $\max U(q_1, \dots, q_n) - C(\sum q_i)$ subject to $\sum q_i = q$. Since cost is fixed, the planner's problem is equivalent to $\max U(q_1, \dots, q_n)$ subject to $\sum q_i = q$. Now consider $W^u(q)$. Since q is the quantity of the monopolist's composite good, $q = D(p) = \sum q_i(p)$, where the outputs $[q_1(p), \dots, q_n(p)]$ maximize utility given $p_i = p$. This means that $D(p)$ solves $\max U(q_1, \dots, q_n)$ subject to $p \sum q_i = p q$, which coincides with the planner's problem. Thus, $W^u(q) = W^*(q)$. Since $W^*(q)$ is the maximum feasible welfare given the constraint $\sum q_i = q$, the first inequality is established.

Consider the second, more novel inequality. Given $q^d < q^u$, it is known that $V(h(q^d)) < V(h(q^u))$. Since q^u is a profit-maximizing output (not necessarily unique) under uniform pricing, $\Pi(q^d) \leq \Pi(q^u)$. Thus, by expression (1), $q^d < q^u$ implies $W^u(q^d) < W^u(q^u)$.

Intuitively, the first inequality reflects the fact that, if the cost function depends only on total output and not on its distribution among goods or markets, then the constraint $\sum q_i = q$ can be interpreted as a particular transformation function, one with marginal transformation rates of unity. Uniform pricing reflects these marginal rates of transformation. Thus, a uniform-price equilibrium will maximize welfare for the given level of total output, while discriminatory

prices generally will not. This is just the same logic that underlies the first welfare theorem.

The second inequality is where the revealed preference argument comes in. It shows that—regardless of the shape of the cost function—welfare under uniform pricing is higher at a profit-maximizing output q^u than at any lower output q^d . For more intuition, express welfare under uniform pricing as total valuation minus total cost: $W^u(q) = B(q) - C(q)$, where B is the integral under the demand curve from 0 to q . Since q^u maximizes profit, moving from a lower output q^d to q^u must increase revenue by at least as much as cost: $\Delta R \geq \Delta C$. Since increasing quantity demanded from q^d to q^u would require lowering price, total valuation would increase by more than revenue: $\Delta B > p^u(q^u - q^d) > \Delta R$. Therefore, $\Delta W = \Delta B - \Delta C > \Delta R - \Delta C \geq 0$, so welfare must increase if, under uniform pricing, output is raised to a profit-maximizing level. Correspondingly, Figure 1 shows welfare at q^u to be higher than at any lower output.

Note that if marginal cost is decreasing and the comparison is of two *arbitrary* outputs, both below the efficient level, then one cannot be sure that welfare will be higher at the higher output. When the cost function is concave, welfare—value minus cost—need not be concave everywhere (even though value is concave) and therefore need not be single-peaked. It is because the higher output represents a profit maximum that one can be sure that welfare there is higher.

I conclude with two remarks about the policy relevance of the analysis. First, the welfare result rests on the assumption that demand curves faced by the monopolist generate adequate measures of welfare. This condition can fail, for example, when the monopolist is selling to distorted intermediate-good markets rather than to final consumers. Consider an input monopolist selling at a uniform price to several unrelated intermediate-good industries. Suppose that in equilibrium the proportional price-cost markups are different in the various industries due to different degrees of competition (rather than different demand elasticities). Then, allocating a given quantity of the in-

put through uniform pricing does not maximize welfare for that input quantity; lower input prices should be charged to the industries with the higher markups. If price discrimination by the input monopolist results in such a pattern, then welfare can be higher under discrimination even if the total quantity of the input is lower. (Such desirable discrimination might be profit-maximizing for the monopolist if, for instance, those industries with the higher markups also have greater ability to substitute in production away from the monopolist's input.) That is, price discrimination by the input monopolist could help counteract the downstream distortions. This is a standard second-best ambiguity.

The second remark concerns the information needed for my result and for those of Schmalensee (1981) and Varian (1985) to provide useful welfare tests in practice (assuming that areas under demand curves do accurately reflect welfare). What must the policymaker know in order to infer that welfare is lower under discrimination if output is observed to be lower? My proposition requires the policymaker to be confident that the monopolist knows demand and cost and that the output observed under uniform pricing, q^u , is profit-maximizing.

Schmalensee (1981) and Varian (1985) require only that marginal cost at q^u be less than price (q^u need not be profit-maximizing, because of the monopolist's imperfect knowledge about cost and demand), provided the policymaker knows also that marginal cost is nondecreasing at lower outputs. Thus, more information is required for the monopolist but less for the policymaker: the policymaker must know only that the monopolist possesses the requisite information needed to maximize profit under uniform pricing.

REFERENCES

- Deaton, Angus and Muellbauer, John, *Economics and Consumer Behavior*, Cambridge: Cambridge University Press, 1980.
- Pigou, A. C., *The Economics of Welfare*, 1st Ed., London: Macmillan, 1920.
- Schmalensee, Richard, "Output and Welfare Implications of Monopolistic Third-Degree Price Discrimination," *American Economic Review*, March 1981, 71, 242-7.
- Varian, Hal R., *Microeconomic Analysis*, 2nd Ed., New York: Norton, 1984.
- , "Price Discrimination and Social Welfare," *American Economic Review*, September 1985, 75, 870-5.

The Measurement of International Trade Related to Multinational Companies

By F. STEB HIPPLE*

Multinational companies (MNCs) are involved in a significant share of U.S. international trade. According to the Bureau of Economic Analysis, U.S.-based multinational companies were associated with 80 percent of export trade and 40 percent of import trade in 1984 (Obie G. Whichard, 1988, p. 87). These aggregate trade shares, however, do not necessarily measure the level of U.S. trade that is uniquely related to MNC operations. This paper will present and analyze four alternate concepts of the definition of MNC-related trade.

Multinational companies evolve from domestic firms that expand their international activities beyond importing and exporting. These firms make equity investments abroad and acquire foreign subsidiaries and affiliates. The emergence of multinational companies has facilitated the evolution of the United States to a more "open" economy where international trade is a major and growing element in production and consumption.

The main source of data and analysis on the activities of multinational companies is the U.S. Bureau of Economic Analysis (BEA), part of the Department of Commerce. The bureau defines the MNC as being a firm based in one country (the "parent") with at least a 10 percent equity interest in a firm located in a second country (the "affiliate"). Information on these firms is provided through periodic benchmark surveys of direct foreign investment and annual reports on MNC operations (Whichard, p. 86). The amount and quality of BEA data on multinational firms have

increased as MNCs have become more significant in U.S. international trade and investment. The set of historical data, however, presents several major problems in addition to the definition of what constitutes MNC-related trade.

First, the detailed data on affiliate operations are not comparable. The data on the foreign affiliates of U.S.-based multinationals focus on the majority-owned affiliates (50 percent equity interest or more). The data on the U.S. affiliates of foreign-based MNCs focus on all affiliates (10 percent equity interest or more). Second, data are available on the parent firms of U.S. multinational companies; no data are available on the parent units of foreign multinationals.

Third, the very important benchmark surveys have been conducted at irregular intervals, and U.S. and foreign MNCs have never been surveyed during the same year. The final problem concerns the timing of the data. The MNCs covered in the benchmark surveys and annual samples have always reported on a fiscal-year basis. This means that comparisons with calendar-year data are never exact, although such comparisons are routinely performed (Whichard, 1988, p. 87).

This paper will examine four different definitions of the trade role of MNCs, including the BEA definition. For comparability, the paper will develop trade data for foreign multinational firms and for U.S. and foreign firms combined. The analytical procedure used in this paper combines a trade matrix and flow-of-funds approach to structure the trading activities of multinational companies. The trade data for U.S. and foreign multinational companies are based on information in the benchmark surveys of foreign direct investment.

*Department of Economics and Finance, College of Business, East Tennessee State University, Johnson City, TN 37614.

I. Multinational Companies and Trade Flows

The trade flows that link the United States and other countries are composed of many individual transactions. These transactions involve buyers and sellers, domestic companies and foreign companies, and the domestic and foreign units of multinational companies. A trade matrix can effectively show the interlinkages among these different types of trade transactors (Hipple, 1982, 1989).

Following the approach of the Bureau of Economic Analysis, divide the world into the United States (U.S.) and the rest of the world (ROW). Assume three types of trade transactors in the world: parent companies *p*, affiliated foreign companies *a*, and other companies *o*. The parent firms in the United States control the affiliates in the rest of the world. The ROW parent firms control the affiliates in the United States. The "other" group of companies are domestic firms that engage in international trade.

The "domestic" affiliates of MNCs in the U.S./ROW world will be divided between the parent and other groups. The majority-owned, domestic affiliates will be counted by the BEA as part of the parent group, while minority-owned, domestic affiliates (10 percent equity interest or more) will fall into the other group of companies. For example, a majority-owned French affiliate of a German-based MNC will be counted as part of the ROW parent; a minority-owned Dutch affiliate of the same MNC will be counted as part of the ROW other group.

Three sets of trade transactors are located in the United States: (1) the parent firms of U.S.-based MNCs, (2) the affiliates of foreign-based MNCs, and (3) other U.S. companies. Three sets of trade transactors are located in the rest of the world: (1) the parent firms of foreign-based MNCs, (2) the affiliates of U.S.-based MNCs, and (3) other ROW companies.

Total U.S. exports *X* may be shown as

$$(1) \quad X = X_p + X_a + X_o.$$

The terms *X_p*, *X_a*, and *X_o* represent exports sold by U.S. parent firms, the affiliates

of foreign-based MNCs, and other U.S. companies, respectively.

The destination of U.S. exports can be shown as

$$(2a) \quad X_p = X_{pp} + X_{pa} + X_{po}$$

$$(2b) \quad X_a = X_{ap} + X_{aa} + X_{ao}$$

$$(2c) \quad X_o = X_{op} + X_{oa} + X_{oo}.$$

The first lowercase letter represents the U.S.-located seller, and the second lowercase letter represents the ROW-located buyer. For example, the term *X_{po}* is the value of exports sold by U.S. parent companies to the other group of foreign (ROW) companies.

Similarly, total U.S. imports *M* can be shown as

$$(3) \quad M = M_p + M_a + M_o.$$

The terms *M_p*, *M_a*, and *M_o* represent imports purchased by U.S. parent firms, the affiliates of foreign-based MNCs, and other U.S. companies, respectively.

The source of U.S. imports can be shown as

$$(4a) \quad M_p = M_{pp} + M_{pa} + M_{po}$$

$$(4b) \quad M_a = M_{ap} + M_{aa} + M_{ao}$$

$$(4c) \quad M_o = M_{op} + M_{oa} + M_{oo}.$$

The first lowercase letter represents the U.S.-located buyer, and the second lowercase letter represents the ROW-located seller. For example, the term *M_{po}* is the value of imports bought by U.S. parent companies from the "other" group of foreign (ROW) companies.

Certain categories of trade transactors have a special interest. The *pa* and *ap* categories represent intrafirm transactions by American-based and foreign-based MNCs, respectively. All other transactions represent "arms-length" relationships. The *pp*, *aa*, and *oo* categories represent overlaps (all "arms-length") among the three types of trade transactors.

II. Four Definitions of MNC-Related Trade

Overall, there are nine categories of paired buyers and sellers in the matrix equation systems in (2) and (4). The four definitions of trade related to multinational companies may be shown as different sets of the nine paired categories.

1. Parent and Affiliate—Under this definition, MNC-related trade includes all trade transactions where the parent firm or the affiliate participates as a buyer and/or seller. In terms of the matrix equations, it would involve all the eight categories with p and/or a . Under this nearly all-inclusive definition, only the so category of trade transactions is excluded. This definition of MNC-related trade is used by the Bureau of Economic Analysis to calculate the level of merchandise trade associated with U.S. multinational companies (Whichard, p. 87).

2. Parent Alone—All trade transactions in which the parent firm participates as a buyer or seller are considered as MNC-related trade. In terms of the matrix equations, it would include the five categories with p . This definition has never received any official use by the Bureau of Economic Analysis, but the data have been published for U.S. multinational companies in the last benchmark survey of U.S. direct investment abroad (U.S. Bureau of Economic Analysis [BEA], 1985, pp. 152, 155).

3. Affiliate Alone—This definition focuses on the trading activities of the affiliated firms. All trade transactions in which the affiliate company participates as a buyer or seller are counted as MNC-related trade. In terms of the matrix equations, it would include the five categories with a . The act of investing in another country creates the multinational company, and the trade activities of the affiliate distinguish the MNC from other companies involved in trade transactions (Hipple, 1982, 1989). A major advantage of this definition is that trade data are available for the affiliates of both U.S. and ROW multinational firms.

4. Intrafirm Shipments—Under this definition, only transactions between the parent and the affiliate are considered as

MNC-related trade. In terms of the matrix equations, it would include only the two categories with pa and ap . This is the most narrow definition of MNC-related trade, and it focuses on the distinction between "arms-length" and "intrafirm" transactions. The transactions between a parent and affiliate are not set by market forces or valued at market prices but represent the production and distribution operations of the vertically integrated multinational company. The values of the transactions are transfer prices set for internal MNC accounting purposes (Hipple, 1982, 1989). As in the previous definition, trade data are available for both U.S. and ROW companies.

III. U.S. Merchandise Trade Flows in 1982

Table 1 is a flow-of-funds table based on the matrix equation system and shows U.S. merchandise trade data for 1982. Trade transactors located in the United States form the row headings; trade transactors located in the rest of the world form the column headings. The trade flows are displayed in "cells" containing U.S. exports, imports, and the resulting trade balance. The cell labels identify the paired categories of trade transactors in equations (2) and (4), and the row and column totals.

The trade data are based on the benchmark surveys of foreign direct investment conducted by the U.S. Bureau of Economic Analysis. The benchmark surveys collect, among other information, detailed data on the exports and imports of U.S. and foreign multinational companies. As described above, more data are available on the operations of U.S. multinationals.

The trade flows shown in Table 1 are taken from two different benchmark surveys. A survey of foreign direct investment in the United States was conducted for 1980, and a survey of U.S. direct investment abroad was conducted for 1982 (BEA, 1983, 1985). These are sometimes referred to as "outward investment" and "reverse investment," respectively. In the tables, the 1980 data from the reverse investment survey have been rescaled to 1982 trade levels. In

TABLE 1—MNC-RELATED TRADE FLOWS IN 1982 U.S. EXPORTS AND IMPORTS
(BILLIONS OF DOLLARS)

Firms Located in the United States	Firms Located in the Rest of the World			Total Trade
	ROW Parent Firms	U.S. Affiliate Firms	Other ROW Firms	
<u>U.S. Parent Firms</u>	(<i>pp</i>)	(<i>pa</i>)	(<i>po</i>)	(<i>pT</i>)
Exports	\$70.803	\$46.559	\$35.863	\$153.225
Imports	65.615	41.598	3.748	110.961
Balance	5.189	4.961	32.114	42.264
<u>ROW Affiliate Firms</u>	(<i>ap</i>)	(<i>aa</i>)	(<i>ao</i>)	(<i>aT</i>)
Exports	\$20.186	\$2.097	\$27.932	\$50.215
Imports	47.621	1.849	27.318	76.788
Balance	-27.435	0.249	0.613	-26.573
<u>Other U.S. Firms</u>	(<i>op</i>)	(<i>oa</i>)	(<i>oo</i>)	(<i>oT</i>)
Exports	\$0.459	\$8.062	\$0.232	\$8.753
Imports	45.637	7.959	2.607	56.203
Balance	-45.178	0.102	-2.375	-47.450
<u>Total Trade</u>	(<i>Tp</i>)	(<i>Ta</i>)	(<i>To</i>)	(<i>TT</i>)
Exports	\$91.448	\$56.718	\$64.027	\$212.193
Imports	158.373	51.406	33.673	243.952
Balance	-67.424	5.312	30.353	-31.759

Note: The terms in parentheses identify cells in the trade matrix on the basis of equations (2) and (4). A *T* is a total of the row or column. The figures in bold type are taken from the BEA benchmark surveys of direct foreign investment or other official sources. All other figures are derived under various assumptions. See the Appendix for the sources and derivation of each cell.

this manner, the trading activities of both U.S. and foreign MNCs can be compared and measured.

In Table 1, six cells are in bold type, and ten cells are in regular type. The data in bold type are taken from the benchmark surveys and other official sources. The data in regular type are derived under various assumptions. The Appendix describes the derivation of the trade data shown in each cell.

The trading activities of multinational affiliates are shown by a row and a column of cells forming a "cross" in the flow-of-funds matrix (Hipple, 1989). The row of cells *ap*, *aa*, *ao*, and *aT* shows the trade activities of the U.S.-located affiliates of foreign-based MNCs. Similarly, the column of cells *pa*, *aa*, *oa*, and *Ta* shows the trade activities of the ROW-located affiliates of

U.S.-based MNCs. Data cell *aa* shows the overlap in the trade activities of the two groups of companies, which must be taken into account to avoid double counting.

The affiliate "cross" and cell *TT* contain five of the six cells in bold type since the benchmark surveys focus primarily on affiliate operations. The affiliate data cells in regular type (*aa*, *ao*, and *oa*) are a simple allocation of residuals based upon a constant market share assumption.

The additional cell in bold type (cell *pT*) provides the total imports and exports associated with U.S. parent firms. The imports and exports with affiliates are known (cell *pa*), but there is no rule to allocate the residual into the two corner cells (*pp* and *po*). A constant market share assumption was used to derive the missing cells in the affiliate "cross" since data were available on

TABLE 2—U.S. TRADE ORIGINATED BY MULTINATIONAL COMPANIES IN 1982
(VALUES IN BILLIONS OF DOLLARS; SHARES IN PERCENT)

	Definition of MNC-Related Trade			
	DEF 1 Parent and Affiliate	DEF 2 Parent Alone	DEF 3 Affiliate Alone	DEF 4 Intrafirm Shipments
United States MNCs				
Export Value	\$163.384	\$153.225	\$56.718	\$46.559
Import Value	120.769	110.961	51.406	41.598
Balance	42.615	42.264	5.312	4.961
Export Share	77.0 percent	72.2 percent	26.7 percent	21.9 percent
Import Share	49.5	45.5	21.1	17.1
Overall Share	62.3	57.9	23.7	19.3
Rest of World MNCs				
Export Value	\$121.477	\$91.448	\$50.215	\$20.186
Import Value	188.040	158.873	76.788	47.621
Balance	-66.562	-67.424	-26.573	-27.435
Export Share	57.2 percent	43.1 percent	23.7 percent	9.5 percent
Import Share	77.1	65.1	31.5	19.5
Overall Share	67.9	54.9	27.8	14.9
Combined MNCs				
Export Value	\$211.961	\$173.870	\$104.836	\$66.745
Import Value	241.345	204.219	126.345	89.219
Balance	-29.384	-30.349	-21.510	-22.474
Export Share	99.9 percent	81.9 percent	49.4 percent	31.5 percent
Import Share	98.9	83.7	51.8	36.6
Overall Share	99.4	82.9	50.7	34.2
Total U.S. Trade				
Export Value	\$212.193			
Import Value	243.952			
Balance	-31.759			

Note: The export and import values under each definition are calculated from the trade matrix cells in Table 1. See the Appendix for the calculation formulas. The export and import trade shares are calculated by dividing the trade values by total U.S. exports and imports. The overall share is a trade-weighted average of the export and import shares.

the market shares of both U.S. and ROW affiliate companies. However, no equivalent information exists on the market shares of parent companies. Data exist for the U.S. parents but not for the ROW parents.

A "heroic" assumption is needed to fill all the data cells in Table 1. The only data on market shares for MNC parents refer to U.S.-based companies. The "heroic" assumption is that ROW parents play the same role in ROW trade as U.S. parents play in U.S. trade. That is, the foreign parent companies have the same trade share of ROW imports and exports as U.S. parents have of U.S. imports and exports. Two points must be noted. First, the trade shares refer to nonaffiliate trade, and second, U.S.

exports (imports) are the ROW imports (exports). With this assumption, the residual from cells pT and pa can be allocated into cells pp and po . A similar procedure fills cells op and oo . Then cells Tp and To can be filled by summing the columns.

IV. Trade Related to Multinational Companies

Table 2 shows the values and shares of MNC-related trade under the four definitions. Data are shown for U.S.-based MNCs, foreign-based MNCs, and the two groups of multinational companies combined. The derivation formulas are shown in the Appendix. Two analytical issues are of interest. First, what is the relationship between MNC

trade and the trade balance? Second, what is the relationship between MNC trade and overall trade levels?

As seen in Table 2, the United States recorded a \$31.8 billion deficit in merchandise trade in 1982. Under definition 1 (the "official" BEA concept), the trading activities of combined U.S. and ROW multinational companies were associated with a \$29.4 billion deficit. This reflects the \$42.6 billion surplus from the activities of U.S. multinationals, the \$66.6 billion deficit from foreign MNCs, and a \$5.4 billion deficit adjustment to reflect the overlapping trade (cells *pp* and *aa*). Under definition 2, these figures are nearly identical.

It is an easy step to credit MNC-trading activities as the cause of the deficit. However, as the share data show, combined MNCs accounted for 99.4 percent of U.S. trade under definition 1 and 82.9 percent under definition 2. The combined MNC trade deficit under definition 1 is simply the overall U.S. merchandise trade deficit. The similarities of the figures under definition 2 and the 82.9 percent trade share lead to a similar conclusion. If the trade role of MNCs is broadly defined, then nearly all merchandise trade transactions will be counted.

Under definitions 3 and 4, MNC-related trade is restricted to a narrower range of merchandise trade transactions (Hipple, 1989). Combined affiliate trade (definition 3) resulted in a \$21.5 billion deficit in 1982, while accounting for 50.7 percent of merchandise trade. This reflects a modest \$5.3 billion surplus from the activities of U.S. affiliates, a \$26.6 billion deficit from ROW affiliates, and a \$0.2 billion deficit adjustment due to overlapping trade (data cell *aa*). Combined intrafirm shipments (definition 4) showed a \$22.5 billion deficit and a 34.2 percent share. The shipments of U.S. multinationals had a \$5.0 billion surplus that was offset by the \$27.4 billion deficit of foreign MNCs. (There is no trade overlap under definition 4). The similarities of these figures under definitions 3 and 4 is striking.

Intrafirm transactions (definition 4) are nested within affiliate transactions (definition 3). The figures under definition 3 are the result of intrafirm shipments plus some

additional transactions. These additional affiliate transactions are "arms-length" and, as shown in Table 2, tend to be equal on the export and import sides. The affiliates of U.S. multinationals have about \$10 billion in additional exports and imports; the affiliates of ROW multinationals have about \$30 billion. Thus, the deficit in 1982 from affiliate trade (definition 3) appears to be the same deficit as from intrafirm trade (definition 4).

These results suggest some findings and some questions. First, benchmark survey data are needed for the activities of ROW parent companies in U.S. merchandise trade. The role of multinational companies cannot be precisely measured without this missing component. The assumption used here is "heroic" but may not be too far from the mark. The combined MNC trade shares in the range of 90 percent under definition 1 and 80 percent under definition 2 seem reasonable given the known levels of trade associated with U.S.-based MNCs.

Second, the usefulness of the BEA definition of MNC-related trade (definition 1) is questionable. There is a near identity between MNC-related trade and total merchandise trade. In this context, separate identification of MNC-related trade provides little information since the multinationals account for nearly all of the trade. The same comments would apply to definition 2.

Third, definitions 3 and 4 usefully analyze MNC-related trade as an important component of total merchandise trade. In 1982, affiliate trade was about one-half of U.S. merchandise trade; intrafirm trade was about one-third. Shifts in the levels and shares of these types of transactions can provide insights into changes in U.S. trade performance (Hipple, 1989). Affiliate trade and intrafirm trade provide information on the changing structure of U.S. trade transactions and the motives underlying trade activities.

APPENDIX

The data shown in the tables have been developed from the benchmark surveys of foreign direct investment conducted periodically by the U.S. Bureau of

Economic Analysis. Benchmark surveys of U.S. direct investment abroad have been conducted for the years 1966, 1977, and 1982. Benchmark surveys of foreign direct investment in the United States have been conducted for the years 1974 and 1980. The foreign direct investment surveys are now on a seven-year cycle. The next benchmark survey of foreign investment in the United States will cover 1987, while the next benchmark survey of U.S. direct investment abroad will cover 1989. Given the normal data processing lags, the information from this last set of benchmark surveys will not be available until 1992.

Table 1 shows U.S. merchandise trade flows for 1982 and is based on the benchmark surveys of 1980 and 1982. The trade data from the 1980 survey of foreign direct investment in the United States have been rescaled to 1982 levels. The trade figures in bold type are taken directly from the benchmark surveys (or other government sources), while all other data in these tables are derived under various assumptions. The derivation of individual data cells in Table 1 is discussed below.

Table 2 shows the trade originated by multinational companies under the four different definitions. The trade values and trade shares are based on the trade flows in Table 1. The calculation formulas for the export and import values are shown below. The trade shares are the export and import values divided by total U.S. exports and imports. The "overall shares" shown in the table are a trade-weighted average of the export and import shares

Derivation of Data Cells in Table 1

(*pp*) The residual of ($pT - pa$) must be allocated into cells *pp* and *po*. It is assumed that the foreign MNC parents play the same role in ROW trade as U.S. parents play in U.S. trade. Thus, the role of U.S. parents in U.S. exports (imports) is replicated by ROW parents in ROW exports (imports) that are U.S. imports (exports). In cell *pp*, exports are $(pTX - paX)(pTM / (TTM - aTX))$, and imports are $(pTM - paM)(pTX / (TTX - aTX))$, where *pTX* is exports in cell *pT*, etc.

(*pa*) (BEA, 1985, pp. 129, 133).

(*po*) Residual of $pT - (pp + pa)$.

(*pT*) (BEA, 1985, pp. 152, 155).

(*ap*) Adjustment factors are used to convert 1980 data to a 1982 basis. For exports, the factor is 0.962, which is the ratio of 1982 merchandise exports to 1980 exports. For imports, the factor is 1.013, which is the ratio of 1982 merchandise imports to 1980 imports (U.S. Bureau of the Census, 1980, p. 11, 1982, p. 9; BEA, 1983, p. 141).

(*aa*) The trade flow data for cells *pa*, *ap*, *aT*, *Ta*, and *TT* are known. The residual amounts ($Ta - pa$) and ($aT - ap$) must be allocated into cells *aa* and *ao*, respectively. The unallocated level of trade within the matrix is $TT - (pa + ap)$. ROW affiliates account for $(Ta - pa)$; thus, the ratio $(Ta - pa) / (TT - (pa + ap))$ is used to allocate ($aT - ap$) into cell *aa*; or, U.S. affiliates account for ($aT - ap$), and the ratio $(aT - ap) / (TT - (pa + ap))$ is used to

allocate ($aT - ap$) into cell *aa*. Either procedure will result in identical estimates for the trade flows in cell *aa*.

(*ao*) Residual of $aT - (ap + aa)$.

(*aT*) Adjustment factors are used to convert 1980 data to a 1982 basis. See the discussion under cell *ap*. (BEA, 1983, p. 141).

(*op*) The residual of ($oT - oa$) must be allocated into cells *op* and *oo*. See the discussion under cell *pp*.

(*oa*) Residual of $Ta - (pa + aa)$.

(*oo*) Residual of $oT - (op + oa)$.

(*oT*) Residual of $TT - (pT + aT)$.

(*TP*) Sum of $pp + ap + op$.

(*Ta*) (BEA, 1985, pp. 127, 131).

(*To*) Sum of $po + ao + oo$.

(*TT*) (Census, 1982, p. 9).

Calculation of Trade Values for Table 2

The export and import trade values shown in Table 2 are calculated from the trade matrix cells in Table 1. DEF 1 is all transactions where a parent or affiliate participates as a buyer or seller. DEF 2 is all transactions where the parent is the buyer or seller. DEF 3 is all transactions where the affiliate participates as a buyer or seller. DEF 4 is limited to intrafirm transactions between the parent and affiliate.

The cells *pp* and *aa* represent an overlap between the trading activities of U.S. and foreign MNCs. Therefore, the export and import values of combined MNCs must be adjusted for the overlap. The cell formulas are as follows.

United States MNCs:

$$\text{DEF 1} = pp + pa + po + aa + oa$$

$$\text{DEF 2} = pp + pa + po$$

$$\text{DEF 3} = pa + aa + oa$$

$$\text{DEF 4} = pa$$

Rest of World MNCs:

$$\text{DEF 1} = pp + ap + op + aa + ao$$

$$\text{DEF 2} = pp + ap + op$$

$$\text{DEF 3} = ap + aa + ao$$

$$\text{DEF 4} = ap$$

Combined MNCs:

$$\begin{aligned} \text{DEF 1} = & pp + pa + po + ap + aa \\ & + ao + op + oa \end{aligned}$$

$$\text{DEF 2} = pp + pa + po + ap + op$$

$$\text{DEF 3} = pa + aa + oa + ap + ao$$

$$\text{DEF 4} = pa + ap.$$

REFERENCES

- Hipple, F. Steb, *The Role of Multinational Firms in U.S. International Trade*. Washington: U.S. Department of Commerce, International Trade Administration, July 1982.
- , "The Changing Role of Multinational Corporations in U.S. International Trade," in H. Peter Gray, ed., *The Modern International Environment*, Greenwich, CT: JAI Press, 1989, 65–80.
- Whichard, Obie G., "U.S. Multinational Companies: Operations in 1986," *Survey of Current Business*, June 1988, 68, 85–96.
- U.S. Bureau of Economic Analysis, *Foreign Direct Investment in the United States, 1980*, Washington: USGPO, October 1983.
- , *U.S. Direct Investment Abroad: 1982 Benchmark Survey Data*, Washington: USGPO, December 1985.
- U.S. Bureau of the Census, *FT990*, December 1980.
- , *FT990*, December 1982.

The Indirect and Direct Substitution Effects

By MASAO OGAKI*

The classification of two goods as substitutes or complements by the sign of the substitution term defined by John R. Hicks (1939) intimately involves the relation of each of the two goods to other goods, as was emphasized by Paul A. Samuelson (1974) and Masazo Sono (1961) among others. Hence, Hicks's definition may lead to a counterintuitive classification when a third good possesses a strong influence. This defect of Hicks's definition motivated Samuelson to propose an alternative definition for substitutes and complements. The purpose of this note is not to propose an alternative classification, but to characterize the effect on the substitution term from a specified third good. Actually, the present note analyzes this effect from multiple goods. This task is important because Hicks's definition is most frequently used in spite of the deficiency.

As discussed in Section I, it is possible to give an intuitive argument as to how the classifications of two goods are affected by a third good. The main goal for my characterization of the effect from a third good given in Section II is to provide the intuitive argument with precise and quantitative content. The effect is characterized by a further decomposition of the substitution term: given a third good, the substitution term will be decomposed into what I call the direct and the indirect substitution terms, the former being free from the effect of the third good,

and the latter characterizing the effect. This characterization enables one to verify whether the given third good causes two goods to be substitutes (complements) when the two goods would be complements (substitutes) if the effect of the third good were eliminated. Section III gives an empirical illustration. Previously (Ogaki, 1989), I applied the concept of the indirect and direct substitution effects to theoretical work in international financial economics.

I. An Intuitive Argument

Samuelson (1974 p. 1255) offered an illuminating example:

...sometimes I like tea and cream... I also sometimes take cream with my coffee. Before you agree that cream is therefore a complement to both tea and coffee, I should mention that I take much less cream in my cup of coffee than I do in my cup of tea. Therefore, a reduction in the price of coffee may reduce my demand for cream, which is an odd thing to happen between so-called complements.

Though Samuelson treats the uncompensated price change here, it is obvious that this example is also applicable to the compensated price change. Thus, coffee and cream may be classified as substitutes rather than complements in Hicks's definition.

This example can be explained as follows. Suppose that a compensated price reduction in the price of coffee is experienced by a consumer. Then there exist two kinds of effects which work in opposite directions on the demand for cream. One kind of effect works directly: since the consumer tends to consume coffee and cream together, the demand for cream is increased. The other kind of effect works indirectly via the demand for tea: he now demands less tea, since coffee and tea are substitutes, and less

*Department of Economics, University of Rochester, Rochester, NY 14627. I thank Lars Hansen, Lionel McKenzie, Jose Schenkman, participants of the Second Buffalo-Cornell-Rochester Conference, and two anonymous referees for helpful comments. I am also grateful to Mahmoud El-Gamal, John Heaton, Dean Lillard, Barbara Mace, Jonathan Ostry, Rangarhan Sundram, and especially Sherwin Rosen for suggestions which improved the exposition and to Kenjiro Hirayama, Noboru Kyotaki, Robert Lucas, and Kiminori Matsuyama for conversations which motivated this work.

consumption of tea leads to less demand for cream. We may call the former the direct substitution effect and the latter the indirect substitution effect. Though coffee and cream are direct complements, coffee and cream are indirect substitutes with respect to tea, as the argument above shows. If the indirect substitution effect is greater than the direct substitution effect in absolute value, coffee and cream are substitutes in Hicks's sense.

In the example, coffee and cream are indirect substitutes, because cream is a complement of a substitute of coffee, namely tea. Similar intuitive argument can be employed to show that a substitute of a substitute is an indirect complement and that a complement of a complement is an indirect complement.

II. Definition and Properties of the Indirect and Direct Substitution Effects

In order to give the intuitive argument in the last section a precise content, I propose a definition of the direct and indirect substitution terms. Suppose there is a compensated price change in coffee, and consider the change in demand for cream. In order to remove the effect of a third good, say tea, consider the change in demand for cream when the consumption of tea is kept constant. This change is the direct substitution effect between coffee and cream with respect to tea. The difference between Hicks's substitution effect and the direct substitution effect may be called the indirect substitution effect.¹ The indirect substitution effect characterizes the effect of the specified third good on the substitution effect. The usefulness of the decomposition depends on the two properties given below. Since the direct substitution effect is nothing but the substitution effect under "straight" or "specific commodity" rationing, properties of both the direct and the indirect substitution effects are closely related with results in the literature of rationing.

¹The direct and indirect substitution effects are defined with respect to a specified third good. However, I omit the phrase "with respect to..." when the third good is clear by context.

A. General Preferences

Consider a consumer facing n goods. When the consumer is unconstrained, the expenditure function takes the form

$$(1) \quad m(p_a, p_b, u) \\ = \inf_{x_a, x_b} [p'_a x_a + p'_b x_b; v(x_a, x_b) > u]$$

where $x_a = (x_1, \dots, x_k)$, $x_b = (x_{k+1}, \dots, x_n)$, $p_a = (p_1, \dots, p_k)$, $p_b = (p_{k+1}, \dots, p_n)$, and $v(x_a, x_b)$ is a utility function. When the consumer is constrained to consume \bar{x}_b of x_b , an expenditure function may be defined as

$$(2) \quad \tilde{m}(\bar{x}_b, p_a, u) \\ = \inf_{x_a} [p'_a x_a; v(x_a, \bar{x}_b) \geq u].$$

It is assumed that $m(p_a, p_b, u)$ is twice continuously differentiable with respect to (p_a, p_b) in a nonempty set T of \mathbb{R}^{n+1} , so that the infimum for $m(p_a, p_b, u)$ is uniquely attained by $\partial m / \partial p_i = x_i^c(p_a, p_b, u)$ for $i = 1, \dots, n$. Here $x_i^c(p_a, p_b, u)$ is compensated (or Hicksian) demand function for the i th good. It is also assumed that $\tilde{m}(\bar{x}_b, p_a, u)$ evaluated at

$$\bar{x}_b = x_b^c(p_a, p_b, u) \\ = [x_{k+1}^c(p_a, p_b, u), \dots, x_n^c(p_a, p_b, u)]$$

is twice continuously differentiable with respect to (\bar{x}_b, p_a) in T . This assumption implies, among the other things, that the infimum for $\tilde{m}(x_b^c(p_a, p_b, u), p_a, u)$ is uniquely attained by $\partial \tilde{m} / \partial p_i = \tilde{x}_i^c(\bar{x}_b, p_a, u)$ for $i = 1, 2, \dots, k$ in T . Note that the compensated constrained demand function for the i th good, $\tilde{x}_i^c(\bar{x}_b, p_a, u)$, does not depend on p_b . Let

$$(3) \quad S_{ij}(p_a, p_b, u) = \partial x_i^c / \partial p_j(p_a, p_b, u) \\ (i, j = 1, \dots, n)$$

be the substitution term.² Define the direct substitution term with respect to the $(k+1)$ th, ..., $(n-1)$ th, and n th goods by

$$(4) \quad S_{ij}^d(p_a, p_b, u) \\ = \partial \tilde{x}_i^c / \partial p_j \cdot x_b^c(p_a, p_b, u), p_a, u) \\ (i, j = 1, \dots, k).$$

It should be noted that the partial derivative is evaluated at $\bar{x}_b = x_b^c(p_a, p_b, u)$. Otherwise, it would not be possible to interpret S_{ij}^d and S_{ij}^i (defined below) as a decomposition of S_{ij} . The indirect substitution term with respect to the $(k+1)$ th, ..., $(n-1)$ th, and n th goods is defined by

$$(5) \quad S_{ij}^i(p_a, p_b, u) \\ = S_{ij}(p_a, p_b, u) - S_{ij}^d(p_a, p_b, u) \\ (i, j = 1, \dots, k).$$

Let $S^d = [S_{ij}^d]_{i,j=1,\dots,k}$ be the matrix of direct and $S^i = [S_{ij}^i]_{i,j=1,\dots,k}$ be the matrix of indirect substitution terms.

It is easily seen that S^d is symmetric and negative semidefinite because $S_{ij}^d = \partial^2 \bar{m} / \partial p_i \partial p_j$. The fact that S^i is negative semidefinite follows from the general properties of restricted expenditure functions and is a special case of a Le Châtelier-type result proved by J. P. Neary and K. W. S. Roberts (1980 p. 3c).

Property 1: For any $(p_a, p_b, u) \in T$, the direct and indirect substitution matrices, S^d and S^i , are symmetric and negative semidefinite.

Because of the symmetry property, the following terminology is legitimate. If the direct (indirect) substitution term between two goods is positive, the goods are called direct (indirect) substitutes; if the term is negative,

they are called direct (indirect) complements.

Let $S_{ib} = [S_{ij}]_{j=k+1,\dots,n}$ be an $(n-k) \times 1$ vector and let $S_{bb} = [S_{ij}]_{i,j=k+1,\dots,n}$ be an $(n-k) \times (n-k)$ matrix of substitution terms.

Property 2: For any $(p_a, p_b, u) \in T$, $S_{ij}^i = S'_{ib}(S^{-1})_{bb}S_{ib}$ for $1 \leq i \leq k$ and $1 \leq j \leq k$ if S_{bb} is nonsingular.

PROOF:³

Differentiating the identity

$$(6) \quad x_i^c(p_a, p_b, u) \\ \equiv \tilde{x}_i^c(x_b^c(p_a, p_b, u), p_a, u) \\ (i = 1, \dots, k)$$

with respect to p_j ($j = 1, \dots, k$) yields

$$(7) \quad S_{ij} = \sum_{q=k+1}^n S_{qj} \frac{\partial \tilde{x}_i^c}{\partial \bar{x}_q} + S_{ij}^d \\ (j = 1, \dots, k).$$

Differentiating (6) with respect to p_j ($j = k+1, \dots, n$) yields

$$(8) \quad S_{ij} = \sum_{q=k+1}^n S_{qj} \frac{\partial \tilde{x}_i^c}{\partial \bar{x}_q} \\ (j = k+1, \dots, n).$$

Combining (5), (7), and (8) completes the proof.

In the special case where there is only one third good ($k+1 = n$), equation (8) is nothing but James Tobin and H. S. Houthakker's (1950-51) equation 3.8, and equation (7) is their equation 5.2.

Property 2 shows that the indirect substitution term between coffee and cream with respect to tea equals the substitution term

²For convenience, S_{ij} is defined as function of (p_a, p_b, u) . On substituting the indirect utility function to u , S_{ij} can be regarded as a function of the prices and income.

³An anonymous referee suggested this proof, which is much simpler than the proof given in the earlier versions of the present note. The proof is an application of Robert A. Pollak's (1969) theory of conditional demand functions.

between coffee and tea multiplied by the substitution term between tea and cream divided by the own substitution term of tea. This property may be called the sign property of the indirect substitution effect: since the own substitution term of tea is always negative, the sign of the indirect substitution term between tea and coffee is determined by those of the substitution term between coffee and tea and the substitution term between tea and cream. It follows from the sign property that a complement of a substitute is an indirect substitute, that a substitute of a substitute is an indirect complement, and that a complement of a complement is an indirect complement. Thus, the definition formalizes the intuitive argument given in the last section. Property 2 also makes it possible to calculate the direct and the indirect substitution effects from only the knowledge of substitution terms. Thus, whenever an empirical researcher obtains a strange result about the substitute-complement relationship of two goods, he or she can check whether some third good renders the two goods substitutes or complements through the indirect substitution effect.

Note that the Property 2 also holds in the elasticity form. Let $e_{ij} = S_{ij}p_j/x_i^c$ be the compensated cross-price elasticity of the j th price on the i th good. Define $e_{ij}^d = S_{ij}^d p_j/x_i^c$, $e_{ij}^i = S_{ij}^i p_j/x_i^c$ as direct and indirect cross-price elasticities, respectively. Then,

$$(9) \quad e_{ij} = e_{ij}^d + e_{ij}^i$$

and by (7), (8), and (9),

$$(10) \quad e_{ij}^i = e'_{bj}(e_{bb}^{-1})e_{ib}$$

where

$$e_{bj} = [e_{ij}]_{i=k+1, \dots, n}$$

$$e_{ib} = [e_{ij}]_{j=k+1, \dots, n}$$

are $(n-k) \times 1$ vectors and $e_{bb} = [e_{ij}]_{i,j=k+1, \dots, n}$ is an $(n-k) \times (n-k)$ matrix.

B. Separable Preferences

Sono (1961) defined the concept of the proper substitution effect in the case of weak separability in preferences. It is of interest to see the relation between the direct substitution effect and the proper substitution effect under separability. Assume that x_a is weakly separable from the goods in x_b , and let $\bar{v}(x_a) = v(x_a, \bar{x}_b)$. Then $v(x_a, x_b) = f(\bar{v}(x_a), x_b)$ for some function f . Suppose that $x_a^p(p_a, u) = [x_1^p(p_a, u), \dots, x_k^p(p_a, u)]$ solves the problem of minimizing $p_a'x_a$ subject to $\bar{v}(x_a) \geq u$. Then, Sono's proper substitution term is $\partial x_i^p / \partial p_j$ for $i, j = 1, \dots, k$. Obviously, $x_i^p(p_a, u) \equiv \bar{x}_i^c(p_a, u)$ must hold. Hence, Sono's proper substitution term is equal to the direct substitution term with respect to x_b under separability. Sono showed that the proper substitution term could have a different sign from S_{ij} , which Sono called the general substitution term. Hence, the indirect substitution effect can dominate the direct substitution term even under separability.

III. An Empirical Illustration

In order to illustrate the use of our results, they are applied to estimates of cross-price elasticities given in Angus S. Deaton (1974). Deaton estimated $c_{ij} = w_i e_{ij}$, using maximum-likelihood methods for British data from 1900 to 1970, where w_i is the budget share of the i th good. We use the budget share w_i given in Deaton (1974 p. 360) to convert estimates of c_{ij} reported in his table III for the symmetric version of the Rotterdam model into elasticities, e_{ij} . Once e_{ij} 's are obtained, indirect and direct cross-elasticities with respect to any third good can be calculated from relations (9) and (10).

Deaton's point estimate of the cross-price elasticity of food and entertainment (con-

TABLE 1—INDIRECT AND DIRECT CROSS-PRICE ELASTICITIES FOR FOOD AND ENTERTAINMENT WITH RESPECT TO ALTERNATIVE THIRD GOODS

Item	e_{ij}^I	SE ^a	e_{ij}^D	SE ^a
Footwear and clothing	-0.063	0.053	0.014	0.079
Housing	-0.005	0.032	-0.044	0.021
Fuel	-0.001	0.044	-0.048	0.018
Drink and tobacco	-0.003	0.033	-0.046	0.018
Travel and communication	-0.002	0.052	-0.047	0.023
Other goods	0.001	0.047	-0.050	0.018
Other services	-0.001	0.065	-0.049	0.019

Note: Estimates reported in table III of Deaton (1974) were used to calculate indirect and direct cross-price elasticities for consumption of food when the price of entertainment changes.

^aThe standard errors reported were calculated with the assumption that the correlations between Deaton's estimates were zero.

sisting of books and magazines, newspapers, and other entertainment) indicates that these two goods are complements. This might be counterintuitive, because there does not seem to exist strong reason to believe that consumption of food and books should be increased simultaneously. In Table 1, indirect and direct cross-price elasticities for food and entertainment are reported for alternative choices of the third good. A mean-value approximation (the delta method) can be used to calculate standard errors of the estimates reported in Table 1. This calculation requires the covariance matrix of estimates of cross-price elasticities. Since Deaton did not report the covariance between different estimates, I assumed that the covariance was zero in the calculation. Hence, the standard errors reported in Table 1 are approximations.

The hypothesis that food and entertainment are direct substitutes can be rejected at the 5-percent level for all potential third goods examined, except for footwear and clothing (clothing for short). The point estimate of the direct-substitution term suggests that food and entertainment are direct substitutes with respect to clothing. Thus, the hypothesis that the classification is due to the indirect effect from the third good, clothing, cannot be rejected. This is because clothing is estimated to be a strong substitute both for food and for entertainment.

REFERENCES

- Deaton, Angus S., "The Analysis of Consumer Demand in the United Kingdom, 1900-1970," *Econometrica*, March 1974, 42, 341-67.
- Hicks, John R., *Value and Capital*, Oxford: Clarendon Press, 1939.
- Neary, J. P. and Roberts, K. W. S., "The Theory of Household Behavior under Rationing," *European Economic Review*, January 1980, 13, 25-42.
- Ogaki, Masao, "Demand for Foreign Bonds and the Term Structure of Interest Rates," unpublished manuscript, April 1989.
- Pollak, Robert A., "Conditional Demand Functions and Consumption Theory," *Quarterly Journal of Economics*, February 1969, 83, 60-78.
- Samuelson, Paul A., "Complementarity: An Essay on the 40th Anniversary of the Hicks-Allen Revolution in Demand Theory," *Journal of Economic Literature*, December 1974, 12, 1255-89.
- Sono, Masazo, "The Effect of Price Changes on the Demand and Supply of Separable Goods," *International Economic Review*, September 1961, 2, 239-75.
- Tobin, James and Houthakker, H. S., "The Effects of Rationing on Demand Elasticities," *Review of Economic Studies*, 1950-51, 18(3), 140-53.

Auction Institutional Design: Theory and Behavior of Simultaneous Multiple-Unit Generalizations of the Dutch and English Auctions

By KEVIN A. MCCABE, STEPHEN J. RASSENTI, AND VERNON L. SMITH*

Historically, English and Dutch auctions have been used for the exchange of single objects such as works of art or single lots of a good such as produce, fish, or cut flowers. Where these institutions have been used for the exchange of multiple units, such as the Australian wool auction (using English rules), successive lots of the good are sometimes sold sequentially at auction. In some, but not all, instances this is because the goods are not identical, even though the various lots may be close substitutes (see Penny Burns, 1985). Where the goods are accepted universally as being homogeneous, as in the securities markets, multiple units are often commonly auctioned simultaneously. In the securities industry, orders are batched for simultaneous execution in multiple-unit auctions in what are referred to as "call markets"; that is, the security is "called" for auction at a particular point in time. This type of market is used on the stock exchanges of Austria, Belgium, France, Germany, and Israel. Some of these are verbal, and some are sealed bid auctions.

Although the U.S. organized exchanges are predominantly continuous rather than call markets (except that call markets are used each day to open trading in each listed security), there is a growing number of exceptions such as the proliferation, since 1984, of Auction Preferred Stock (Goldman, Sachs and Co., October 1984) and Money Market Preferred Stock (Lehman Brothers, July 1984). We now have Dutch Auction Rate Transferable Securities, called DARTS, Stated Rate Auction Preferred

Stock, or STRAPS, and many more. After the initial subscription offering of this type of security, the market is called every 49 days to reset the preferred dividend rate using a multiple-unit auction. The exchange of shares and the dividend determination is based on the array of stated dividend rates at which existing holders and potential new holders are willing to sell and/or buy corresponding quantities. The dividend rate and exchange of shares every 49 days is executed using the uniform price or competitive sealed bid mechanism (Vernon L. Smith et al., 1980). The discussion to follow will be confined to this sealed bid form of the call market.

Call markets provide temporal consolidation of trade orders or other forms of expressing the desire to buy and sell. By comparison with continuous trading, call markets offer both advantages and disadvantages (Robert A. Schwartz, 1988 pp. 442-6). The cited advantages include low cost of operating the exchange; information aggregation and presumed pricing efficiency; price stability; individual trades, which are thought to have a small impact on price; reduced price uncertainty; and, finally, nondiscriminatory pricing. However, there are offsetting disadvantages: (1) the market is inaccessible except at the time of call; (2) no bid, offer, contract, or price information is available until the results of the call are announced; and (3) there is transaction uncertainty because a submitted bid (offer) may be too low (high) to execute inside the supply-demand cross. These conditions are only partially alleviated if there is a secondary market between calls.

These disadvantages may be significant. In September 1988, the *Wall Street Journal* published an article on the failure of a call market for the auction rate preferred stock

*Economic Science Laboratory, University of Arizona, Tucson, Arizona. This material is based upon work supported by the National Science Foundation under grant no. SES-8320121.

of Kroger Co. The article lists several other call market failures in addition to the Kroger issue.

We are currently researching the theoretical and behavioral properties of a number of new proposed institutions that represent alternatives to the sealed bid "call" auction. These are exercises in institutional design. Previous experimental research has established that the two-sided uniform price sealed bid-offer auction is less efficient and yields less competitive prices than does the continuous double auction. In fact, no institution studied to date is more efficient than the latter. Our objective is to explore a variety of new institutional designs in search of an institution at least as efficient as the continuous double auction, that has all of the advantages claimed for call markets, and that corrects for disadvantages (2) and (3) cited above. That is, we seek new trading institutions suitable for call market application in which traders receive some form of information feedback enabling them to make desired adjustments and to reduce uncertainty as to whether they will be able to transact.

I. Institutional Design: Two New Mechanisms for Call Market Exchange

In this paper, we compare multiple-unit generalizations of the Dutch and English auction institutions when there is an inelastic supply of four homogeneous units of a good and ten prospective buyers. Since the offer price is adjusted automatically by clock in both of the institutions we study, we refer to them as the Dutch clock and English clock mechanisms. It is common information that the buyers desire at most one unit of the good and that the buyers' reservation values have been drawn randomly from a uniform distribution over 1-cent discrete increments from $[0, 224]$.¹

¹Subjects for the experiment were recruited from the undergraduate population at the University of Arizona. They were paid three dollars at the beginning of each experiment as an incentive to show up. At the end of each experiment subjects were paid their salient earnings in U.S. dollars.

This study is motivated by the seminal work of William Vickrey (1961, 1962, 1976) and the objective of overcoming some of the disadvantages of the sealed bid form of call market organization. We also extend the experimental research on single- and multiple-unit auctions. For a single-unit environment, Vicki M. Coppinger, Vernon L. Smith, and Jon A. Titus (1980) find that their experimental data are consistent with some of the predictions made by William Vickrey (1961). Thus, mean prices are statistically the same for English and Dutch auctions; the variance in Dutch prices is less than the variance in English prices. James C. Cox, Bruce Roberson, and Vernon L. Smith (1982) generalize the single-unit Vickrey model to allow heterogeneous risk aversion. This generalization explains the lower efficiency of the Dutch auction when compared to the English auction.

Our proposed multiple-unit Dutch and English clock mechanisms are shown to have the same theoretical properties as the corresponding single-unit Dutch clock and English oral auctions. We find that most of the qualitative differences between Dutch and English single-unit auctions extend to nondiscriminatory multiple-unit versions of these institutions.

Each experiment consisted of 22 auctions, all of which used either the Dutch or English institutional rules.² Each auction was run as follows. First, each subject was assigned a private resale value on a Plato terminal. After each subject had been assigned a value at random, the auction was begun. The auction took place at the front of the room with one of the experimenters acting as auctioneer. When the auction was over, the common market price for the good and the actual winners were announced. This information was then recorded on each subject's Plato terminal. In addition, each

²Once subjects had all arrived, they were given a set of written instructions for the auction institution used that day. Subjects were asked to read the instructions and then listen to a brief example presented by one of the experimenters. After the instructions were finished and any questions answered, the experiment was begun.

subject's profit was displayed privately on the terminal. After each subject had examined the auction outcome and his private profit, the next auction period was begun.

If a subject did not win a unit, he made zero profit; otherwise, a subject's profit was the difference between his resale value and the market price. Subjects could make negative profits in any period, but total profit in the experiment was constrained to be non-negative. This constraint was never binding. Subjects were paid cash equal to three times their total profit (rounded up to the nearest quarter) at the end of the experiment; to all subjects this was common information.

II. English Clock: Institution and Theory

The English clock is set initially at zero, and all buyers who wish to buy at this price are asked to raise their identification cards. The clock then begins to increment by 5 cents (until five subjects are active and then by 1 cent). As long as buyers wish to stay in the auction, they keep their cards raised. Once a buyer lowers his card, he is out of the auction. The clock continues to rise until only four buyers remain. As soon as the fifth buyer drops out of the auction, the clock is stopped and the market price paid by the remaining four winners is the price on the clock. (See Ralph Cassiday [1967] for discussion of an experimental English clock for single object exchange.)

The theory of the single-object English auction generalizes naturally to the multiple-unit English clock. Each buyer has a dominant strategy to stay in the auction as long as the clock price is below his or her resale value.³ When the clock rises to a buyer's resale value, he or she is motivated to drop out of the auction. The predicted equilibrium price is the fifth highest resale value. The auction is Pareto optimal since

the players with the four highest resale values are predicted to each win a unit of the good.

III. Dutch Clock: Institution and Theory

In our experiments, the Dutch clock is set at 225, and all buyers who wish to buy at this price are asked to raise their ID cards and keep them up. The clock then begins to decrement by 5 cents until three subjects have their cards raised, then by 1 cent. The clock continues to fall until four buyers have their cards raised. At this point, the clock is stopped and the market price paid by the four winners is the price on the clock.

Consider a theoretical version of this institution in which an analogue clock drops continuously. Let $v_i \in [0, 1]$ be the value to player i of buying a unit. We assume that v_i 's are chosen independently from a continuous uniform distribution; the value v_i is private information; there are N buyers in the auction; there are $M < N$ units to be sold; and the buyers are risk-neutral.

Denote by b_{ij} player i 's reservation price for raising his card when j units (M -number raised cards) remain to be sold. Vickrey has shown that

$$(1) \quad b_{i1}^* = \frac{N-1}{N} v_i$$

is a Nash equilibrium bid function when $M = 1$.

Let $\alpha_M, \dots, \alpha_1$ be an arbitrary set of fractions that satisfy $1 \geq \alpha_M \geq \alpha_{M-1} \geq \dots \geq \alpha_1 = (N-M)/(N-M+1)$. Then the following is a Nash equilibrium bidding strategy for $M > 1$.

$$(2) \quad b_{ij}^* = \alpha_j v_i.$$

Equation (2) defines a class of Nash equilibrium bidding strategies since $\alpha_M, \dots, \alpha_2$ can be any order-preserving sequence of proportions as long as $\alpha_2 \geq (N-M)/(N-M+1) = \alpha_1$.

For convenience, let $v_1 > v_2 > \dots > v_N$. By assumption, the probability of a tied

³This theory is based on the standard assumption that values are drawn from a density function and therefore ties have zero probability. In fact, of course, draws are from a discrete mass function, but the probability of a tie is only 1 in 50,625.

value is zero. Note that, by following (2), players $1, \dots, M-1$ will each receive a unit. Player M will determine the price at

$$(3) \quad P^* = \frac{N-M}{N-M+1} v_M = \alpha_1 v_M$$

and also receives a unit. The remaining players $M+1, \dots, N$ will not receive a unit.

Given that all other players bid according to (2), can any individual do better by bidding other than (2)? Assume k units remain to be sold. First, we consider the possibility of ever bidding less than (2). If $k=1$, we have arrived at the one-unit endgame, and Vickrey has shown that bidding $\alpha_1 v$ is optimal. When $k > 1$, no player who would not otherwise win before the endgame can help himself by bidding less than $\alpha_1 v$. For any player who would otherwise win before the endgame, bidding less than $\alpha_k v$ cannot improve his price unless he consistently bids low enough to force himself into the endgame. But that never pays off since he would replace a lower-valued player as the price setter and set a higher price $\alpha_1 v$ for himself.

Now let us consider the possibility of bidding more than suggested by (2). An individual with value v is trying to maximize his expected gain E_k from the remainder of the auction. $M-k$ players already have their cards raised, and $N-M+k-1$ others remain competitive in the auction. Let $\alpha_{k+1} w_k$ be the bid tendered by the most recent ($M-k$ th) player to raise his card. Then the values of the remaining competitors are distributed uniformly on $[0, w_k]$. Now let us describe his expected return as a function of his reservation price p_k for raising his card when k units remain to be sold:

$$(4) \quad E_k = \left(\frac{p_k}{\alpha_k w_k} \right)^{N-M+k-1} \times \left(v - \alpha_1 \frac{(N-M+1)}{(N-M+k)} \frac{p_k}{\alpha_k} \right) + \int_{p_k/\alpha_k}^{w_k} E_{k-1}^* \times \frac{(N-M+k-1)(w_{k-1})^{(N-M+k-2)}}{w_k^{(N-M+k-1)}} dw_{k-1}.$$

The first term of (4) represents expected gain from being the next player to raise his card, while the second term represents his expected gain when another card is raised first, forcing him to continue playing in the $k-1$ unit subgame. The elements of (4) can be described as follows. The expression $(p_k/\alpha_k w_k)^{N-M+k-1}$ is the probability that all remaining $N-M+k-1$ players bid less than p_k given that their values are uniform on $[0, w_k]$ and they use the bidding ratio α_k . The expression $\alpha_1[(N-M+1)/(N-M+k)] p_k/\alpha_k$ is the expected bid of the price setter who has the $k-1$ th highest value of the remaining $N-M+k-1$ players whose values are uniform on $[0, p_k/\alpha_k]$. The variable E_{k-1}^* is the maximized expected return for the $k-1$ unit subgame he must play when someone with a value $w_{k-1} > p_k/\alpha_k$ raises his card first. The expression $(N-M+k-1)w_{k-1}^{(N-M+k-2)}/w_k^{(N-M+k-1)}$ is the density function of the first-order statistic of the remaining $N-M+k-1$ players whose values are uniform on $[0, w_k]$.

For notational convenience, we now define a recursive function $R_k(w_k, v)$ as follows:

$$(5) \quad R_1 = 1; R_k = k + \frac{(N-M+k)}{v} \times \int_v^{w_k} R_{k-1} dw_{k-1} \text{ for } k > 1.$$

By solving the first-order condition of (4), $dE_k/dp_k = 0$, for the cases $k=2$ and 3 , we can show $p_k^* = \alpha_k v$, and the following form for E_k^* results:

$$(6) \quad E_k^* = \frac{v^{(N-M+k)}}{(N-M+k)w_k^{(N-M+k-1)}} R_k.$$

Taking the first-order condition for the general case of (4), with (5) and (6) appropriately substituted, we get the following

equation:

$$\begin{aligned}
 (7) \quad 0 = & \frac{(N-M+k-1)}{\alpha_k w_k^{(N-M+k-1)}} \\
 & \times \left(\frac{p_k}{\alpha_k} \right)^{(N-M+k+2)} v \\
 & - \frac{(N-M)}{\alpha_k w_k^{(N-M+k-1)}} \left(\frac{p_k}{\alpha_k} \right)^{(N-M+k-1)} \\
 & - \frac{v^{(N-M+k-1)}}{\alpha_k w_k^{(N-M+k-1)}} \\
 & \times \left((k-1) + \frac{(N-M+k-1)}{v} \right. \\
 & \left. \times \int_v^{p_k/\alpha_k} R_{k-2} dw_{k-2} \right).
 \end{aligned}$$

An obvious root of (7) is $p_k = \alpha_k v$ since the integral term disappears, there is a common denominator, and the numerator becomes $\{(N-M+k-1) - (N-M) - (k-1)\}v^{(N-M+k-1)} = 0$. Thus, by induction, no player can do any better at any stage of the game by bidding more than $\alpha_k v$. Hence (2) is a class of Nash equilibria.

If players are risk averse and can be characterized by the Arrow-Pratt measure of constant relative risk aversion $(1-r_i)$ for bidder i , with r_i in $(0, 1]$ then James C. Cox, Bruce Roberson, and Vernon L. Smith (1982) have shown that the Nash bidding rule for our endgame is $6/6+r_i$ of the resale value.⁴ If the r_i is distinct for each i among the seven remaining bidders, then a buyer other than the one with fourth highest value could win the fourth unit, thus changing subjects' optimal strategies.

⁴This theory is based on the standard assumption that values are drawn from a density function and therefore ties have zero probability. In fact, of course, draws are from a discrete mass function, but the probability of a tie is only 1 in 50,625.

IV. Experimental Data

Figure 1 graphs actual market price minus the predicted price for each of the four clock experiments. In Table 1 we provide comparisons of mean prices and price variances. A paired t -test comparing actual English prices with predicted English prices results in a t value equal to -2.121 , which is not significant at the 1-percent level. In 22 of the 44 auctions the predicted price exactly equaled the actual price, while in an additional 14 auctions actual price was within 1 cent of the predicted price. In all but one case the difference is nonpositive. (In that case the buyer with the fifth highest resale value left the market slightly later than predicted.) These results differ from the single-unit oral English auction, where actual prices tend to be slightly above predicted prices. This is explained by the fact that Vicki M. Coppinger, Vernon L. Smith, and Jon A. Titus (1980) conducted an English oral auction (not an English clock auction) in which bids were advanced from the floor by the bidders. This causes some overbidding, depending upon the size of the increment by which the highest-value bidder advances the penultimate bid. For a study of the serious strategic problems in English multiple-unit auctions that are created when bidders advance the price from the floor, see Kevin A. McCabe, Stephen J. Rassenti, and Vernon L. Smith (1988).

Figure 2 plots the cumulative distributions of actual and theoretical Dutch prices. A paired t -test in Table 1 (with a t equal to 3.963) confirms that actual prices are significantly greater (at the 1-percent level) than predicted prices. This is consistent with any risk-aversion model of subject behavior.

As can be observed in Figure 1, the Dutch clock tends to create much greater variability in price differences when compared to the English clock. However, as predicted by the theory, the variance in the raw observed Dutch prices (858) is less than the variance in the raw observed English prices (926). F tests in Table 1 show that the differences in variances between actual and predicted

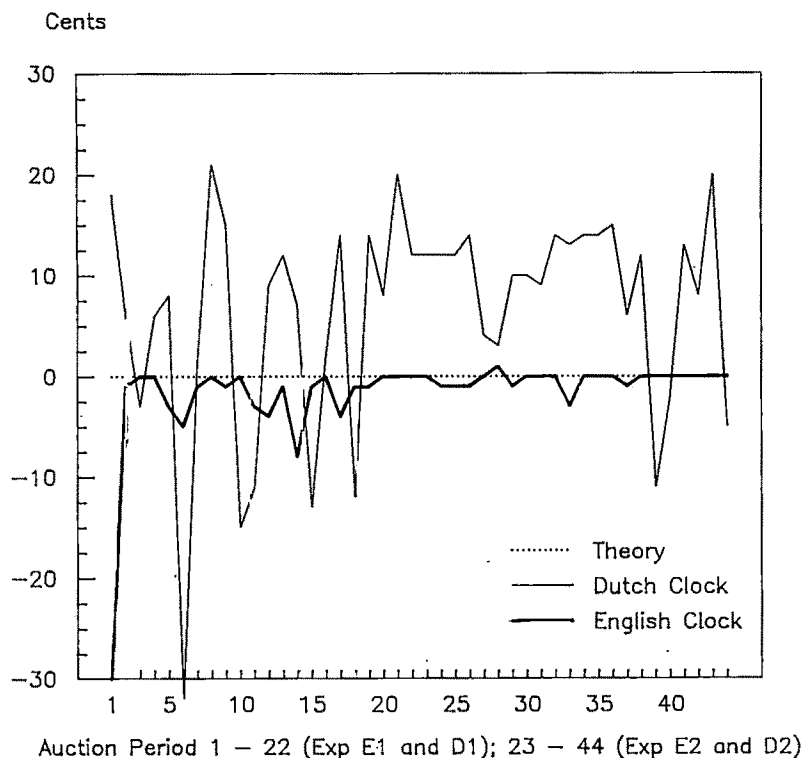


FIGURE 1. GRAPH OF THE DIFFERENCE BETWEEN THE ACTUAL EXPERIMENT PRICE AND THE THEORETICAL PREDICTION IN EACH AUCTION PERIOD. FOR THE DUTCH CLOCK, WE USE THE RISK-NEUTRAL PREDICTION

TABLE 1—COMPARISON OF MEANS AND VARIANCES USING PAIRED *t*-TESTS AND *F* TESTS (43 DEGREES OF FREEDOM)

Series Compared	Means, <i>t</i> Value	Variances, <i>F</i> Value
Dutch Actual and Dutch Theoretical (Conditional on Values)	3.963	1.225
English Actual and English Theoretical (Conditional on Values)	-2.121	1.026
Dutch Actual and English Actual	2.895	1.079

prices are insignificant for both clock auctions.⁵

The mean Dutch price is 119 compared to 111 for the mean English price. This

⁵We calculate an *F* value of 1.225 for actual versus predicted price variances in the Dutch clock experiments and an *F* value of 1.026 for actual versus predicted price variances in the English clock experiments. Neither *F* is significant at the 10-percent level.

difference is significant at the 1-percent level with a paired-test *t* value in Table 1 of 2.895 and is larger than that generally observed in the single-unit oral auctions, due in part, perhaps to our explanation of why the English oral auction price is higher than the English clock price. However, in the last 15 auctions the mean Dutch price was 124 and the mean English price was 120, suggesting some convergence in this measure.

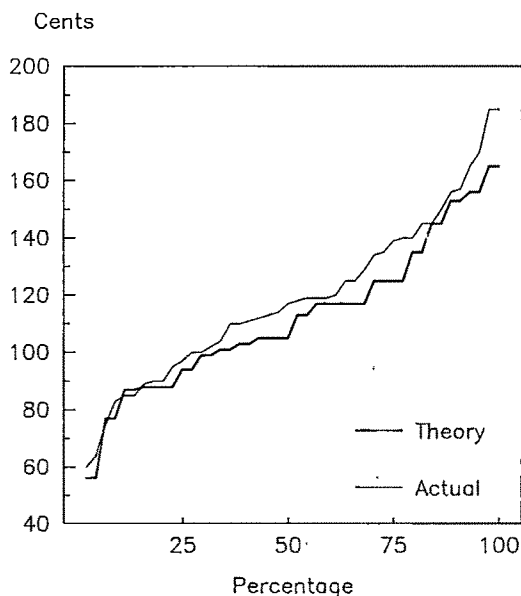


FIGURE 2. CUMULATIVE DISTRIBUTION OF PRICES FOR DUTCH AUCTION EXPERIMENTS, COMPARISON OF ACTUAL VS. PREDICTED DUTCH CLOCK AUCTION FOR 44 AUCTION PERIODS

Efficiency can be measured as actual buyer surplus (the sum of the winners' resale values) divided by the predicted surplus (the sum of the four highest resale values). We find that the English clock is more efficient, with an efficiency of 1.0 in 43 of the 44 auctions.⁶ The one exception is an efficiency of 0.9856 in period 5 of experiment E1. The Dutch clock achieves an efficiency of 1.0 in 28 of the 44 auctions, with an efficiency as low as 0.8661 in period 11 of experiment D1. By the constant-relative-risk-aversion model, this is interpreted as implying that subjects have differing parameter values, r_i .

Finally, we can compare the revenue-generating properties of the two clock auctions. Figure 3 plots the average revenue for each

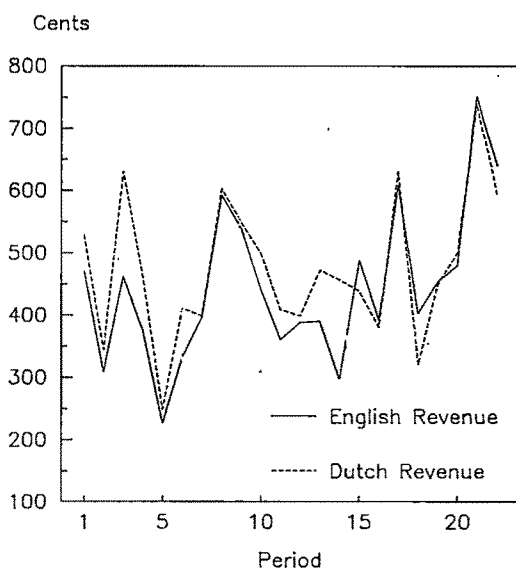


FIGURE 3. AVERAGE REVENUE BY AUCTION PERIOD FOR THE ENGLISH AND DUTCH CLOCKS

clock (two experiments in each clock) for each of the 22 auctions. The Dutch clock generates more revenue than the English clock. Over all 44 auctions, the Dutch clock produces an average revenue of 475 cents, whereas the English clock produces an average revenue of 445 cents.

REFERENCES

- Burns, Penny, "Experience and Decision Making," in V. Smith, ed., *Research in Experimental Economics*, Vol. 3, Greenwich, CT: JAI Press, 1985, 139-58.
- Cassiday, Ralph, *Auctions and Auctioneering*, Los Angeles: University of California Press, 1967.
- Coppinger, Vicki M., Smith, Vernon L. and Titus, Jon A., "Incentives and Behavior in English, Dutch and Sealed-Bid Auctions," *Economic Inquiry*, January 1980, 18, 1-22.
- Cox, James C., Roberson, Bruce and Smith, Vernon L., "Theory and Behavior of Single Unit Auctions," in V. Smith, ed., *Research in Experimental Economics*, Vol. 2, Greenwich, CT: JAI Press, 1982, 1-43.
- McCabe, Kevin A., Rassenti, Stephen J. and Smith, Vernon L., "Testing Vickrey's and

⁶A paired t -test between Dutch and English surplus results in a t value equal to -3.193 , which is significant at the 1-percent level. Furthermore a paired t -test comparing the theoretical maximum surplus with the English clock surplus is insignificant with a t value of 1.00 .

- Other Simultaneous Multiple Unit Versions of the English Auction," in V. Smith, ed., *Research in Experimental Economics*, Vol. 4, Greenwich, CT: JAI Press, forthcoming.
- Schwartz, Robert A., *Equity Markets*, New York: Harper & Row, 1988.
- Smith, Vernon L., Williams, Arlington W., Bratton, William and Vannoni, Michael G., "Competitive Market Institutions: Double Auctions Versus Sealed Bid-Offer Auctions," *American Economic Review*, March 1980, 70, 59-77.
- Vickrey, William, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *Journal of Finance*, March 1961, 16, 8-37.
- _____, "Auctions and Bidding Games," in *Recent Advances in Game Theory*, Proceedings of a Conference, Princeton, NJ: Princeton University Press, 1962, 15-27.
- _____, "Auctions, Markets, and Optimal Allocations," in Y. Amihud, ed., *Bidding and Auctioning for Procurement and Allocation*, New York: New York University Press, 1976, 13-20.
- Winkler, Matthew, "Failure of Kroger Co. Issue Is Viewed as Possible Bad Sign for Auction Market," *Wall Street Journal*, September 1988, p. 40.
- Goldman, Sachs and Co., "Auction Preferred Stock," memorandum of October 19, 1984.
- Lehman Brothers, *American Express Company Prospectus*, July 25, 1984.
- _____, "Money Market Preferred Stock," July 1984.

ERRATA

Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records

By JOSHUA D. ANGRIST*

I. Errors in the Figures

Due to a typographer's error, the figures in my article published in *The American Economic Review* (Volume 80, No. 3, June 1990, pp. 313–36) were transposed and incorrectly labelled. The figure on page 315, labelled Figure 1, is actually Figure 3; the figure on page 324, labelled Figure 3, is actually Figure 1. The title and caption for Figure 1 appear under Figure 3, and vice versa. A corrected set of figures is reproduced in this note, and the figures are briefly described here as well as in the original paper.

Men were called for conscription in the Vietnam era draft lottery according to lottery numbers ranging from 1 to 365 which were randomly assigned to dates of birth. Men with lottery numbers below the highest number called for induction are referred to in the paper as "draft-eligible." Figure 1 shows the history of FICA (Social Security) taxable earnings for draft-lottery participants born between 1950 and 1953. For each cohort and race, two lines are drawn: one for draft-eligible men and one for men with lottery numbers that exempted them from the draft.

Figure 2 presents a magnified view of the effect of draft eligibility on earnings. This figure plots the time series of *differences* in earnings by draft-eligibility status. The figure shows no difference in earnings by draft-eligibility status before the year of conscription risk (1970–3 for men born 1950–3), while in subsequent years the earnings of draft-eligible men generally fall be-

low the earnings of men who could not be drafted.

Figure 3 graphs mean W-2 (federal income taxable) earnings in 1981–4 by cohort and lottery number (\bar{y}_{ctj}) against probabilities of veteran status by cohort and lottery number (\hat{p}_{cj}). Earnings are in 1978 dollars. Plotted in the figure are the average (over four years of earnings) residuals from a regression of earnings and probabilities on period (δ_t) and cohort (β_c) effects. Thus, the slope of the regression line drawn through the points corresponds to an estimate of α in

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}$$

which is equation (3) in the paper. Estimates of equation (3) are equivalent to instrumental variables estimates of

$$y_{cti} = \beta_c + \delta_t + s_i\alpha + u_{it}$$

where s_i indicates veteran status and the instruments are dummy variables that indicate lottery number, cohort, and year of earnings. The parameter α represents the effect of veteran status on earnings and is estimated as –2,384 dollars with a standard error of 778 dollars.

II. Error in Footnote 7

The formula given in Footnote 7 for the sampling variance of the Wald estimates reported in Table 3 is incorrect. The correct variance formula is

$$(\hat{p}^e - \hat{p}^n)^{-2} [\Phi + \alpha^2\phi]$$

where Φ is the variance of $\bar{y}^e - \bar{y}^n$ and ϕ is the variance of $\hat{p}^e - \hat{p}^n$. The formula used

*Department of Economics, Littauer Center, Harvard University, Cambridge, MA 02138.

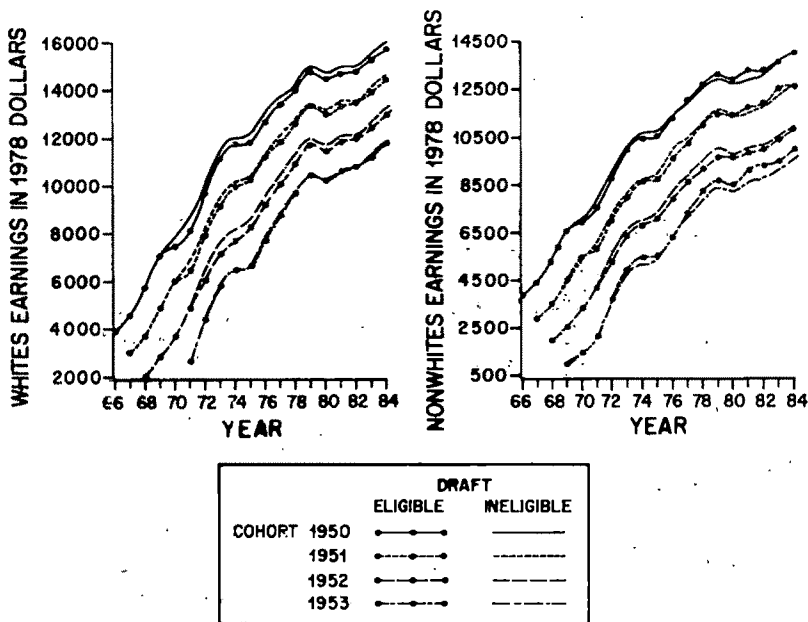


FIGURE 1. SOCIAL SECURITY EARNINGS PROFILES BY DRAFT-ELIGIBILITY STATUS

Notes: The figure plots the history of FICA taxable earnings for the four cohorts born 1950–3. For each cohort, separate lines are drawn for draft-eligible and draft-ineligible men. Plotted points show average real (1978) earnings of working men born in 1953, real earnings + \$3000 for men born in 1950, real earnings + \$2000 for men born in 1951, and real earnings + \$1000 for men born in 1952.

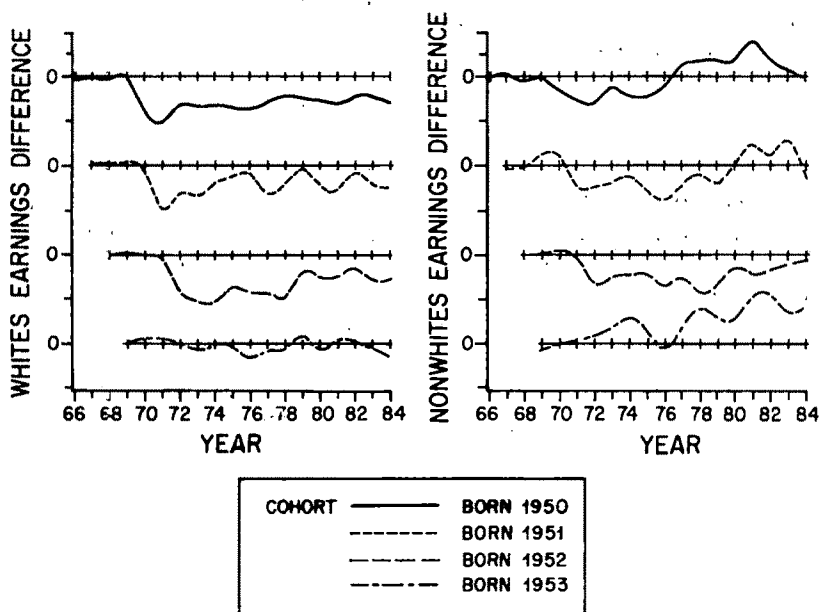


FIGURE 2. THE DIFFERENCE IN EARNINGS BY DRAFT-ELIGIBILITY STATUS

Notes: The figure plots the difference in FICA taxable earnings by draft-eligibility status for the four cohorts born 1950–3. Each tick on the vertical axis represents \$500 real (1978) dollars.

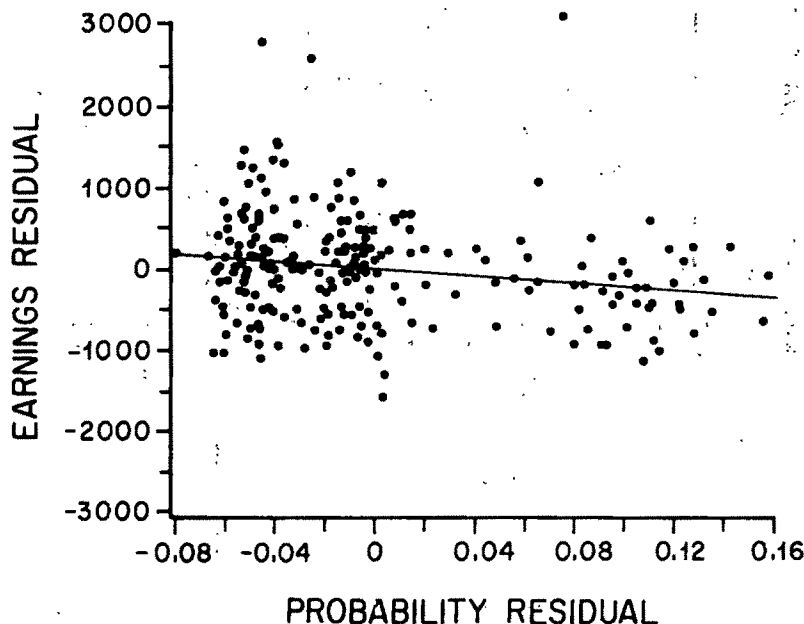


FIGURE 3. EARNINGS AND THE PROBABILITY OF VETERAN STATUS BY LOTTERY NUMBER

Notes: The figure plots mean W-2 compensation in 1981-4 against probabilities of veteran status by cohort and groups of five consecutive lottery numbers for white men born 1950-3. Plotted points consist of the average residuals (over four years of earnings) from regressions on period and cohort effects. The slope of the least-squares regression line drawn through the points is -2,384, with a standard error of 778, and is an estimate of α in the equation

$$\bar{y}_{ctj} = \beta_c + \delta_t + \hat{p}_{cj}\alpha + \bar{u}_{ctj}.$$

in the paper assumes $\alpha = 0$. Replacing α with a consistent estimate in the correct formula gives the variance for the two-sample instrumental variables estimate that is

equivalent to the Wald estimate. As an empirical matter, use of the correct formula raises the standard errors in Table 3 by roughly 10 percent.

ERRATA

On the Optimal Tax Base for Commodity Taxation

By JOHN DOUGLAS WILSON*

The paper, "On the Optimal Tax Base for Commodity Taxation," by John Douglas Wilson (*American Economic Review*, December 1989, Volume 79, No. 5, pp. 1196-1206) contains several typographical errors. Equation (12) on page 1200 should be

$$(12) \quad \gamma: \partial V / \partial \gamma$$

$$+ \lambda [t(\partial b / \partial \gamma)y - dC / d\gamma] = 0.$$

Equation (19) on page 1201, although correct if properly interpreted, should be written in the following form:

$$(19) \quad \frac{\partial MB(\sigma, \gamma, t^*(\sigma, \gamma, R))}{\partial \gamma}$$

*Department of Economics, Ballantine Hall, Indiana University, Bloomington, IN 47405.

$$+ \left[\frac{\partial MB(\sigma, \gamma, t^*(\sigma, \gamma, R))}{\partial t} \times \frac{\partial t^*(\sigma, \gamma, R)}{\partial \gamma} \right] - \frac{\partial MC(\gamma)}{\partial \gamma} \leq 0.$$

The reference to Subsection II.B in the last paragraph on page 1201 should be to Subsection I.B. The last line of equation (A2) on page 1205 should be

$$\times b(1-b)\log(1-t).$$

Finally, the first line of equation (A7) on page 1205 should be

$$(A7) \quad L_{tt} = -\lambda [1 + (1-b)(\sigma - 1)]$$

(i.e., L_{tt} , not L_t).

ERRATUM

The Future of the Income Tax

By JOSEPH A. PECHMAN*

A serious error appeared in Joseph A. Pechman's 1989 Presidential Address, which was published in the March 1990 issue of *The American Economic Review*. The erroneous passage occurs on page 7 at the end of the second full paragraph.

*Joseph Pechman passed away on August 19, 1989. This correction was communicated by Henry Aaron, The Brookings Institution, 1775 Massachusetts Avenue, N.W., Washington, DC 20036.

The original text reads as follows: "Burtless estimates that the Reagan tax and transfer policies increased average annual taxes of men aged 25-54 by no more than 2-4 percent and of women in the same age group by no more than 3.5 percent." The text should read: "Burtless estimates that the Reagan tax and transfer policies increased average annual *labor supply* of men aged 25-54 by no more than 2-4 percent and of women in the same age group by no more than 3.5 percent." (Emphasis added.)

THE AMERICAN ECONOMIC REVIEW

VOLUME LXXX

BOARD OF EDITORS

GEORGE A. AKERLOF
JAMES E. ANDERSON
TIMOTHY F. BRESNAHAN
JOHN Y. CAMPBELL
HENRY S. FARBER
MARJORIE A. FLAVIN
ROBERT P. FLOOD
CLAUDIA D. GOLDIN
JO ANNA GRAY
REUBEN GRONAU
DANIEL S. HAMERMESH
ROBERT J. HODRICK
KEVIN D. HOOVER
KENNETH L. JUDD
JOHN H. KAGEL

JOHN F. KENNAN
DALE T. MORTENSEN
MAURICE OBSTFELD
EDGAR O. OLSEN
JOHN G. RILEY
RICHARD ROLL
THOMAS ROMER
DAVID E. M. SAPPINGTON
KENNETH J. SINGLETON
ROBERT S. SMITH
BARBARA J. SPENCER
RICHARD TRESCH
KENNETH WEST
JOHN D. WILSON
LESLIE YOUNG

EDITOR

ORLEY ASHENFELTER

CO-EDITORS

ROBERT H. HAVEMAN

BENNETT T. McCALLUM

PAUL R. MILGROM

HAL R. VARIAN

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: 209 Nassau Street, Princeton, NJ 08542-4607

Copyright 1990

All Rights Reserved

AMERICAN ECONOMIC ASSOCIATION

CONTENTS OF ARTICLES AND SHORTER PAPERS

J. A. Pechman: The Future of the Income Tax . .	1	G. J. Borjas: Reply	305
K. Rogoff: Equilibrium Political Budget Cycles . .	21	W. A. Bomberger: The Effects of Fiscal Policies When Incomes Are Uncertain: A Contradiction to Ricardian Equivalence: Comment	309
G. Tabellini and A. Alesina: Voting on the Budget Deficit	37	J. D. Angrist: Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records	313
M. R. Garfinkel: Arming as a Strategic Investment in a Cooperative Equilibrium	50	R. N. Stavins and A. B. Jaffe: Unintended Impacts of Public Investments on Private Decisions: The Depletion of Forested Wetlands	337
A. J. Robson: Stackelberg and Marshall	69	W. K. Viscusi and W. N. Evans: Utility Functions That Depend on Health Status: Estimates and Economic Implications	353
W. P. Rogerson: Quality vs. Quantity in Military Procurement	83	R. C. Fair and R. J. Shiller: Comparing Information in Forecasts from Econometric Models	375
P. Bolton and D. Scharfstein: A Theory of Predation Based on Agency Problems in Financial Contracting	93	B. Champ and S. Freeman: Money, Output, and the Nominal National Debt	390
J. Farrell and C. Shapiro: Horizontal Mergers: An Equilibrium Analysis	107	S. G. Cecchetti, P.-S. Lam, and N. C. Mark: Mean Reversion in Equilibrium Asset Prices . .	398
J. A. Ordovery, G. Saloner, and S. C. Salop: Equilibrium Vertical Foreclosure	127	M. B. Canzoneri and C. A. Rogers: Is the European Community an Optimal Currency Area? Optimal Taxation Versus the Cost of Multiple Currencies	419
E. Maskin and D. Newbery: Disadvantageous Oil Tariffs and Dynamic Consistency	143	T. Ito: Foreign Exchange Rate Expectations: Micro Survey Data	434
L. Auernheimer and G. A. Lozada: On the Treatment of Anticipated Shocks in Models of Optimal Control with Rational Expectations: An Economic Interpretation	157	J.-P. Chavas and T. L. Cox: A Non-Parametric Analysis of Productivity: The Case of U.S. and Japanese Manufacturing	450
R. W. Cooper and J. C. Haltiwanger: Inventories and the Propagation of Sectoral Shocks	170	D. S. Scharfstein and J. C. Stein: Herd Behavior and Investment	465
R. D. Sauer and K. B. Laffler: Did the Federal Trade Commission's Advertising Substantiation Program Promote More Credible Advertising?	191	K. Matsuyama: Perfect Equilibria in a Trade Liberalization Game	480
A. Tversky, P. Slovic, and D. Kahneman: The Causes of Preference Reversal	204	D. A. Graham, R. C. Marshall, and J.-F. Richard: Differential Payments Within a Bidder Coalition and the Shapley Value	493
R. W. Cooper, D. V. DeJong, R. Forsythe, and T. W. Ross: Selection Criteria in Coordination Games: Some Experimental Results	218	P. R. Milgrom and J. Roberts: The Economics of Modern Manufacturing: Technology, Strategy, and Organization	511
J. B. Van Huyck, R. C. Battalio, and R. O. Beil: Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure	234	I. Hansson and C. Stuart: Malthusian Selection of Preferences	529
E. Dekel and S. Scotchmer: Collusion Through Insurance: Sharing the Cost of Oil Spill Cleanups	249	L. A. Whittington, J. Alm, and H. E. Peters: Fertility and the Personal Exemption: Implicit Pronatalist Policy in the United States	545
H. P. Young: Progressive Taxation and Equal Sacrifice	253	R. K. Triest: The Relationship Between the Marginal Cost of Public Funds and Marginal Excess Burden	557
K. R. Ihlanfeldt and D. L. Sjoquist: Job Accessibility and Racial Differences in Youth Employment Rates	267	H. Hwang and C.-C. Mai: Effects of Spatial Price Discrimination on Output, Welfare, and Location	567
D. A. Barbezat and J. W. Hughes: Sex Discrimination in Labor Markets: The Role of Statistical Evidence: Comment	277	H. Yoshikawa: On the Equilibrium Yen-Dollar Rate	576
W. E. Even: Comment	287		
P. J. Kuhn: Reply	290		
G. Jasso and M. R. Rosenzweig: Self-Selection and the Earnings of Immigrants: Comment . .	298		

B. L. Benson, M. L. Greenhut, and G. Norman: On the Basing-Point System	584	J. H. Kagel, D. N. MacDonald, and R. C. Battalio: Tests of "Fanning Out" of Indifference Curves: Results from Animal and Human Experiments	912
M. Feldstein and D. W. Elmendorf: Government Debt, Government Spending, and Private Sec- tor Behavior: Revisited: Comment	589	Z. Safra, U. Segal, and A. Spivak: Preference Reversal and Nonexpected Utility Behavior ..	922
F. Modigliani and A. G. Sterling: A Further Com- ment	600	H. J. Kiesling: Economic and Political Founda- tions of Tax Structure: Comment	931
R. C. Kormendi and P. G. McGuire: Reply and Update	604	N. Naqvi: Technological Stagnation, Tenurial Laws, and Adverse Selection: Comment	935
J. L. Stevens: Tobin's q and the Structure-Per- formance Relationship: Comment	618	S. Donnenfeld and L. J. White: Quality Dis- tortion by a Discriminating Monopolist: Com- ment	941
E. Fehr: Cooperation, Harrassment, and Involun- tary Unemployment: Comment	624	S. Rosefelde: Comparative Productivity: Com- ment	946
A. Lindbeck and D. J. Snower: Reply	631	A. Bergson: Reply	955
A. D. Haight: Internal Migration and Urban Em- ployment: Comment	637	D. D. Haddock: On the Basing-Point System: Com- ment	957
I. Henriques: Cooperative and Noncooperative R&D in Duopoly with Spillovers: Comment ..	638	B. L. Benson, M. L. Greenhut, and G. Norman: Reply	963
G. Wright: The Origins of American Industrial Success, 1879-1940	651	R. P. H. Fishe and M. Wohar: The Adjustment of Expectations to a Change in Regime: Com- ment	968
D. Card: Unexpected Inflation, Real Wages, and Employment Determination in Union Con- tracts	669	N. G. Mankiw, J. A. Miron, and D. N. Weil: Reply	977
C. Engel and J. D. Hamilton: Long Swings in the Dollar: Are They in the Data and Do Markets Know It?	689	G. Gennotte and H. Leland: Market Liquidity, Hedging, and Crashes	999
M. P. Keane and D. E. Runkle: Testing the Rati- onality of Price Forecasts: New Evidence from Panel Data	714	L. M. Ausubel: Insider Trading in a Rational Expectations Economy	1022
G. A. Hardouvelis: Margin Requirements, Vol- atility, and the Transitory Component of Stock Prices	736	J.-J. Laffont and J. Tirole: Optimal Bypass and Cream Skimming	1042
G. W. Stadler: Business Cycle Models with En- dogenous Technology	763	D. A. Butz: Durable-Good Monopoly and Best- Price Provisions:	1062
K. Bagwell and R. W. Staiger: A Theory of Man- aged Trade	779	P. S. Segerstrom, T. C. A. Anant, and E. Dinopou- los: A Schumpeterian Model of the Product Life Cycle	1077
G. M. Grossman and E. Helpman: Comparative Advantage and Long-Run Growth	796	M. Dudey: Competition by Choice: The Effect of Consumer Search on Firm Location Deci- sions	1092
W. Darity, Jr.: The Fundamental Determinants of the Terms of Trade Reconsidered: Long- Run and Long-Period Equilibrium	816	J. Dearden, B. W. Ickes, and L. W. Samuelson: To Innovate or Not To Innovate: Incentives and Innovation in Hierarchies	1105
K. Krishna: The Case of the Vanishing Re- venues: Auction Quotas with Monopoly	828	C. Y. C. Chu and H.-W. Koo: Intergenerational Income-Group Mobility and Differential Fer- tility	1125
M. D. Whinston: Tying, Foreclosure, and Exclu- sion	837	M. M. Pitt, M. R. Rosenzweig, and M. N. Hassan: Productivity, Health, and Inequality in the Intrahousehold Distribution of Food in Low- Income Countries	1139
M. Waterson: The Economics of Product Pat- ents	860	H. Holländer: A Social Exchange Approach to Voluntary Cooperation	1157
D. Besanko and D. F. Spulber: Are Treble Dam- ages Natural? Sequential Equilibrium and Pri- vate Antitrust Enforcement	870	M. Dotsey: The Economic Effects of Production Taxes in a Stochastic Growth Model	1168
C. D. Kolstad, T. S. Ulen, and G. V. Johnson: <i>Ex Post</i> Liability for Harm vs. <i>Ex Ante</i> Safety Regulation: Substitutes or Complements? ...	888	M. C. Keeley: Deposit Insurance, Risk, and Mar- ket Power in Banking	1183
Y.-P. Chu and R.-L. Chu: The Subsidence of Preference Reversals in Simplified and Mar- ketlike Experimental Settings: A Note	902		

A. Bar-Ilan: Overdrafts and the Demand for Money	1201	M. Schwartz: Third-Degree Price Discrimination and Output: Generalizing a Welfare Result ..	1259
H. Bohn: Tax Smoothing with Financial Instruments	1217	F. S. Hipple: The Measurement of International Trade Related to Multinational Companies ..	1263
A. F. Daughety: Beneficial Concentration	1231	M. Ogaki: The Indirect and Direct Substitution Effects	1271
D. Levin: Horizontal Mergers: The 50-Percent Benchmark	1238	K. A. McCabe, S. J. Rassenti, and V. L. Smith: Auction Institutional Design: Theory and Behavior of Simultaneous Multiple-Unit Generalizations of the Dutch and English Auctions	1276
P. DeGraba: Input Market Price Discrimination and the Choice of Technology	1246		
B. Nahata, K. Ostaszewski, and P. K. Sahoo: Direction of Price Changes in Third-Degree Price Discrimination	1254		

CONTENTS OF THE PAPERS AND PROCEEDINGS

Richard T. Ely Lecture

- D. S. Landes:** Why Are We So Rich and They So Poor? 1

Economic Education at the High School Level

- W. Becker, W. Greene, and S. Rosen:** Research on High School Economic Education 14

The New Economics of Personnel: Rationale or Rationalization?

- D. J. Aron:** Firm Organization and the New Economic Approach to Personnel Behavior ... 23

- D. M. Gordon:** Who Bosses Whom? The Intensity of Supervision and the Discipline of Labor 28

- S. M. Jacoby and D. J. B. Mitchell:** Sticky Stories: Economic Explanations of Employment and Wage Rigidity 33

Aggregate Asset Pricing

- A. B. Abel:** Asset Prices under Habit Formation and the Catching Up with the Joneses 38

- J. Y. Campbell:** Measuring the Persistence of Expected Returns 43

- S. G. Cecchetti and N. C. Mark:** Evaluating Empirical Tests of Asset Pricing Models: Alternative Interpretations 48

Stock Market Volatility

- W. F. M. De Bondt and R. M. Thaler:** Do Security Analysts Overreact? 52

- R. J. Schiller:** Market Volatility and Investor Behavior 58

- D. M. Cutler, J. M. Poterba, and L. H. Summers:** Speculative Dynamics and the Role of Feedback Traders 63

The Economics and Politics of Budget Control

- D. E. Wildasin:** Budgetary Pressures in the EEC: A Fiscal Federalism Perspective 69

- E. M. Gramlich:** U.S. Federal Budget Deficits and Gramm-Rudman-Hollings 75

- R. P. Inman:** Public Debts and Fiscal Politics: How To Decide 81

The "New" Growth Theory

- G. M. Grossman and E. Helpman:** Trade, Innovation, and Growth 86

- R. E. Lucas, Jr.:** Why Doesn't Capital Flow from Rich to Poor Countries? 92

- P. M. Romer:** Are Nonconvexities Important for Understanding Growth? 97

Lessons for Development from the Experience in Asia

- S. M. Collins:** Lessons from Korean Economic Growth 104

- A. O. Krueger:** Asian Trade and Growth Lessons 108

- T. N. Srinivasan:** External Sector in Development: China and India, 1950-89 113

- Y. C. Park:** Development Lessons from Asia: The Role of Government in South Korea and Taiwan 118

Incentive Effects of Medical Malpractice

- P. M. Danzon, M. V. Pauly, and R. S. Kington:** The Effects of Malpractice Litigation on Physicians' Fees and Incomes 122

- F. A. Sloan:** Experience Rating: Does It Make Sense for Medical Malpractice Insurance?.. 128

New Classical Macroeconomics in the Open Economy

- A. C. Stockman:** International Transmission and Real Business Cycle Models 134

- K. Rogoff:** Bargaining and International Policy Cooperation 139

- R. Dornbusch:** The New Classical Macroeconomics and Stabilization Policy 143

The New Theory of the Firm

- A. Shleifer and R. W. Vishny:** Equilibrium Short Horizons of Investors and Firms 148

- P. R. Milgrom and J. Roberts:** The Efficiency of Equity in Organizational Decision Processes 154

- B. C. Greenwald and J. E. Stiglitz:** Asymmetric Information and the New Theory of the Firm: Financial Constraints and Risk Behavior 160

State and Local Government Finance

- R. Bahl and J. Martinez-Vazquez:** Inflation and the Real Growth of State and Local Government Expenditures 166

- H. F. Ladd:** State Assistance to Local Governments: Changes During the 1980s 171

- E. A. Hanushek and J. M. Quigley:** Commercial Land Use Regulation and Local Government Finance 176

The Rationality of the Foreign Exchange Rate

- J. A. Frankel and K. A. Froot:** Chartists, Fundamentalists, and Trading in the Foreign Exchange Market 181

- R. J. Hodrick:** Volatility in the Foreign Exchange Market and Stock Markets: Is It Excessive?.. 186

R. A. Meese and A. K. Rose: Nonlinear, Nonparametric, Nonessential Exchange Rate Estimation	192	<i>Women's Labor Market Mobility: Evidence from the NLS</i>	
<i>Interaction Between Environmental and Agriculture Policies: Opportunities for Coordination and Limitations for Evaluation</i>		J. A. Klerman and A. Leibowitz: Child Care and Women's Return to Work After Childbirth . .	284
R. E. Just and J. M. Antle: Interactions Between Agricultural and Environmental Policies: A Conceptual Framework	197	M. A. Hill: Intercohort Differences in Women's Labor Market Transitions	289
S. R. Johnson, R. Wolcott, and S. V. Aradhyula: Coordinating Agricultural and Environmental Policies: Opportunities and Trade-offs . .	203	A. Light and M. Ureta: Gender Differences in Wages and Job Turnovers Among Continuously Employed Workers	293
J. Hrubovcak, M. LeBlanc, and J. Miranowski: Limitations in Evaluating Environmental and Agricultural Policy Coordination Benefits . .	208	<i>Wage Trends and the Job Creation Debate</i>	
<i>Bargaining and Price Formation under Incomplete Information: Theories and Experiments</i>		J. S. Leonard and L. Jacobson: Earnings Inequality and Job Turnover	298
V. P. Crawford: Explicit Communication and Bargaining Outcomes	213	B. Bluestone: The Impact of Schooling and Industrial Restructuring on Recent Trends in Wage Equality in the United States	303
A. Schotter: Bad and Good News about the Sealed-Bid Mechanism: Some Experimental Results	220	M. H. Koster: Schooling, Work Experience, and Wage Trends	308
S. R. Williams: The Transition from Bargaining to a Competitive Market	227	<i>Selectivity Bias: New Developments</i>	
<i>The National Research Council's Report on the Status of Black Americans, 1940-85</i>		J. J. Heckman: Varieties of Selection Bias	313
T. D. Boston: A Common Destiny: How Does It Compare to the Classic Studies of Black Life in America?	232	C. F. Manski: Nonparametric Bounds for Treatment Effects	319
R. Farley: Blacks, Hispanics, and White Ethnic Groups: Are Blacks Uniquely Disadvantaged?	237	W. K. Newey, J. L. Powell, and J. R. Walker: Semiparametric Estimation of Selection Models: Some Empirical Results	324
J. J. Heckman: The Central Role of the South in Accounting for the Economic Progress of Black Americans	242	<i>Reviving the Federal Statistical System</i>	
W. A. Darity, Jr.: Racial Inequality in the Managerial Age: An Alternative to the NRC Report	247	J. A. Miron and C. D. Romer: Reviving the Federal Statistical System: The View from Academia	329
<i>The Formation of Economic Values</i>		R. Cole: Reviving the Federal Statistical System: A View from Industry	333
H. Kunreuther and D. Easterling: Are Risk-Benefit Trade-offs Possible in Siting Hazardous Facilities?	252	R. E. Lipsey: Reviving the Federal Statistical System: International Aspects	337
W. K. Viscusi: Sources of Inconsistency in Societal Response to Health Risks	257	J. E. Triplett: Reviving the Federal Statistical System: A View from Within	341
E. Karni and D. Schmeidler: Fixed Preferences and Changing Tastes	262	<i>The Economic History of Technology</i>	
<i>New Developments in Economic Theory</i>		D. C. Mowery: The Development of Industrial Research in U.S. Manufacturing	345
B. Allen: Information as an Economic Commodity	268	J. Mokyr: Punctuated Equilibria and Technological Progress	350
D. Fudenberg and E. Maskin: Evolution and Cooperation in Noisy Repeated Games	274	P. A. David: The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox	355
H. M. Polemarchakis: The Economic Implications of an Incomplete Asset Market	280	<i>Poverty and the Underclass</i>	
		M. Corcoran, R. Gordon, D. Laren, and G. Solon: Effects of Family and Community Background on Economic Status	362
		P. Gottschalk: AFDC Participation Across Generations	367
		M. Tienda: Welfare and Work in Chicago's Inner City	372

Are We Saving Too Little?

- B. P. Bosworth:** International Differences in Saving 377
M. Woodford: Public Debt as Private Liquidity .. 382

Deregulated Airline Markets

- S. A. Morrison and C. Winston:** The Dynamics of Airline Pricing and Competition 389
S. T. Berry: Airport Presence as Product Differentiation 394
S. Borenstein: Airline Mergers, Airport Dominance, and Market Power 400

Strikes, Bargaining, and Arbitration: New Development

- J. Kennan and R. Wilson:** Can Strategic Bargaining Models Explain Collective Bargaining Data? 405
D. Card: Strikes and Bargaining: A Survey of the Recent Empirical Literature 410
O. Ashenfelter and J. Currie: Negotiator Behavior and the Occurrence of Disputes 416

The Industrial Organization of Markets with Differentiated Products

- D. J. Aron and E. P. Lazear:** The Introduction of New Products 421
A. Shepard: Pricing Behavior and Vertical Contracts in Retail Markets 427
D. P. Mann and J. P. Wissink: Money-Back Warranties vs. Replacement Warranties: A Simple Comparison 432

The AEA Commission on Graduate Education in Economics

- W. Lee Hanson:** Educating and Training New Economics Ph.D's: How Good a Job Are We Doing? 437
A. S. Blinder: Discussion 445
C. Golding: Discussion 445
T. P. Schultz: Discussion 447
R. M. Solow: Discussion 448

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

- Alesina, A. 37
 Alm, J. 545
 Anant, T. C. A. 1077
 Angrist, J. D. 313
 Auernheimer, L. 157
 Ausubel, L. M. 1022
 Bagwell, K. 779
 Barbezat, D. A. 277
 Bar-Ilan, A. 1201
 Battalio, R. C. 234, 912
 Beil, R. O. 234
 Benson, B. L. 584, 963
 Bergson, A. 955
 Besanko, D. 870
 Bohn, H. 1217
 Bolton, P. 93
 Bomberger, W. A. 309
 Borjas, G. J. 305
 Butz, D. A. 1062
 Canzoneri, M. B. 419
 Card, D. 669
 Cecchetti, S. G. 398
 Champ, B. 390
 Chavas, J.-P. 450
 Chu, C. Y. C. 1125
 Chu, R.-L. 902
 Chu, Y.-P. 902
 Cooper, R. W. 170, 218
 Cox, T. L. 450
 Darity, W., Jr. 816
 Daughety, A. F. 1231
 Dearden, J. 1105
 DeGraba, P. 1246
 DeJong, D. V. 218
 Dekel, E. 249
 Dinopoulos, E. 1077
 Donnenfeld, S. 941
 Dotsey, M. 1168
 Dudey, M. 1092
 Elmendorf, D. W. 589
 Engel, C. 689
 Evans, W. N. 353
 Even, W. E. 287
 Fair, R. C. 375
 Farrell, J. 107
 Fehr, E. 624
 Feldstein, M. 589
 Fische, R. P. H. 968
 Forsythe, R. 218
 Freeman, S. 390
 Garfinkel, M. R. 50
 Gennotte, G. 999
 Graham, D. A. 493
 Greenhut, M. L. 584, 963
 Grossman, G. M. 796
 Haddock, D. D. 957
 Haight, A. D. 637
 Haltiwanger, J. C. 170
 Hamilton, J. D. 689
 Hansson, I. 529
 Hardouvelis, G. A. 736
 Hassan, Md. N. 1139
 Helpman, E. 796
 Henriques, I. 638
 Hipple, F. S. 1263
 Holländer, H. 1157
 Hughes, J. W. 277
 Hwang, H. 567
 Ickes, B. W. 1105
 Ihlanfeldt, K. R. 267
 Ito, T. 434
 Jaffe, A. B. 337
 Jasso, G. 298
 Johnson, G. V. 888
 Kagel, J. H. 912
 Kahneman, D. 204
 Keane, M. P. 714
 Keeley, M. C. 1183
 Kiesling, H. J. 931
 Koo, H.-W. 1125
 Kolstad, C. D. 888
 Kormendi, R. C. 604
 Krishna, K. 828
 Kuhn, P. 290
 Laffont, J.-J. 1042
 Lam, P.-S. 398
 Leffler, K. B. 191
 Leland, H. 999
 Levin, D. 1238
 Lindbeck, A. 631
 Lozada, G. A. 157
 MacDonald, D. N. 912
 Mai, C.-C. 567
 Mankiw, N. G. 977
 Mark, N. C. 398
 Marshall, R. C. 493
 Maskin, E. 143
 Matsuyama, K. 480
 McCabe, K. A. 1276
 Meguire, P. G. 604
 Milgrom, P. 511
 Miron, J. A. 977
 Modigliani, F. 600
 Nabata, B. 1254
 Naqvi, N. 935
 Newbery, D. 143
 Norman, G. 584, 963
 Ogaki, M. 1271
 Ordovery, J. A. 127
 Ostaszewski, K. 1254
 Pechman, J. A. 1
 Peters, H. E. 545
 Pitt, M. M. 1139
 Rassenti, S. J. 1276
 Richard, J.-F. 493
 Roberts, J. 511
 Robson, A. J. 69
 Rogers, C. A. 419
 Rogerson, W. P. 83
 Rogoff, K. 21
 Rosefield, S. 946
 Rosenzweig, M. R. 298, 1139

Ross, T. W. 218
 Runkle, D. E. 714
 Safra, Z. 922
 Sahoo, P. K. 1254
 Saloner, G. 127
 Salop, S. G. 127
 Samuelson, L. W. 1105
 Sauer, R. D. 191
 Scharfstein, D. 93, 465
 Schwartz, M. 1259
 Scotchmer, S. 249
 Segal, U. 922
 Segerstrom, P. S. 1077
 Shapiro, C. 107
 Shiller, R. J. 375
 Sjoquist, D. L. 267
 Slovic, P. 204
 Smith, V. L. 1276
 Snower, D. J. 631
 Spivak, A. 922
 Spulber, D. F. 870
 Stadler, G. W. 763

Staiger, R. W. 779
 Stavins, R. N. 337
 Stein, J. C. 465
 Sterling, A. G. 600
 Stevens, J. L. 618
 Stuart, C. 529
 Tabellini, G. 37
 Tirole, J. 1042
 Triest, R. K. 557
 Tversky, A. 204
 Ulen, T. S. 888
 Van Huyck, J. B. 234
 Viscusi, W. K. 353
 Waterson, M. 860
 Weil, D. N. 977
 Whinston, M. D. 837
 White, L. J. 941
 Whittington, L. A. 545
 Wohar, M. 968
 Wright, G. 651
 Yoshikawa, H. 576
 Young, H. P. 253

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

Abel, A. B. 96
 Allen, B. 268
 Antle, J. M. 197
 Aradhyula, S. V. 203
 Aron, D. J. 23, 431
 Ashenfelter, O. 416
 Bahl, R. W. 166
 Becker, W. 14
 Berry, S. T. 394
 Blinder, A. S. 445
 Bluestone, B. 303
 Borenstein, S. 400
 Boston, T. D. 232
 Bosworth, B. P. 377
 Campbell, J. Y. 43
 Card, D. E. 410
 Cecchetti, S. G. 48
 Cole, R. 333
 Collins, S. M. 104
 Corcoran, M. 362
 Crawford, V. P. 213
 Currie, J. N. 416
 Cutler, D. M. 63
 Danzon, P. M. 122
 Darity, W. A., Jr. 247
 David, P. A. 355
 De Bondt, W. F. M. 52
 Dornbusch, R. 143
 Easterling, D. 252
 Farley, R. 237
 Frankel, J. A. 181
 Froot, K. A. 181
 Fudenberg, D. D. 274
 Goldin, C. D. 445

Gordon, D. M. 28
 Gordon, R. 362
 Gottschalk, P. 367
 Gramlich, E. M. 75
 Greene, W. 14
 Greenwald, B. C. 160
 Grossman, G. M. 86
 Hansen, W. L. 437
 Hanushek, E. A. 176
 Heckman, J. J. 242, 313
 Helpman, E. 86
 Hill, M. A. 289
 Hodrick, R. J. 186
 Hrubovcak, J. 208
 Inman, R. P. 81
 Jacobson, L. 298
 Jacoby, S. M. 33
 Johnson, S. R. 203
 Just, R. E. 197
 Karni, E. 262
 Kennan, J. F. 405
 Kington, R. S. 122
 Klerman, J. A. 284
 Kosters, M. H. 308
 Krueger, A. O. 108
 Kunreuther, H. 252
 Ladd, H. F. 171
 Landes, D. S. 1
 Laren, D. 362
 Lazear, E. P. 421
 LeBlanc, M. 208
 Leibowitz, A. 284
 Leonard, J. S. 298
 Light, A. L. 293

Lipsey, R. E. 337	Schmeidler, D. 262
Lucas, R. E., Jr. 92	Schotter, A. R. 220
Mann, D. P. 432	Schultz, T. P. 447
Manski, C. F. 319	Shepard, A. 427
Mark, N. C. 48	Shiller, R. J. 58
Martinez-Vazquez, J. 166	Shleifer, A. 148
Maskin, E. 274	Sloan, F. A. 128
Meese, R. A. 192	Solon, G. R. 362
Milgrom, P. R. 154	Solow, R. M. 448
Miranowski, J. 208	Srinivasan, T. N. 113
Miron, J. A. 329	Stiglitz, J. E. 160
Mitchell, D. J. B. 33	Stockman, A. C. 134
Mokyr, J. 350	Summers, L. H. 63
Morrison, S. A. 389	Thaler, R. M. 52
Mowery, D. C. 345	Tienda M. 372
Newey, W. K. 324	Triplett, J. E. 341
Park, Y. C. 118	Ureta, M. 293
Pauly, M. V. 122	Viscusi, W. K. 257
Polemarchakis, H. M. 280	Vishny, R. W. 148
Poterba, J. M. 63	Walker, J. R. 324
Powell, J. L. 324	Wildasin, D. E. 69
Quigley, J. M. 176	Williams, S. R. 227
Roberts, D. J. 154	Wilson, R. B. 405
Rogoff, K. S. 139	Winston, C. D. 389
Romer, C. D. 329	Wissink, J. P. 432
Romer, P. M. 97	Wolcott, R. 203
Rose, A. K. 192	Woodford, M. 382
Rosen, S. 14	



"The speediest and most versatile of the mathematical programming languages."

Barry Simon - PC Magazine

You get more work done when you use the right language.



Fast and Flexible

Now you can discover why scientists, engineers and statisticians agree that GAUSS is the language of choice when you want the most in speed and flexibility from your PC. The built-in

```
Multiple Regression
xxi = inv(x'x);
b = xxi*(x'y);
sse = (y-x*b)'(y-x*b)/(n-rows(b));
siderr = sqrt(diag(sse*xxi));
tvalue = b ./ siderr;
pvalue = cdf(|tvalue|,rows(x));
```

matrix routines are written in assembly language and are faster than even the fastest C or Fortran compilers.

It's not just execution speed that makes GAUSS the better choice. GAUSS is a matrix language, using familiar syntax for expressing mathematical relationships. The result is you will shorten your development time and write less code. GAUSS's programming environment has all the tools you will need to write large programs, including an on-line debugger and an on-line help system that can handle the functions you write.

Expandable

Create your own libraries of functions or call FORTRAN, C, or assembly language routines.

Publication Quality Graphics

The GAUSS system includes 2-D and 3-D publication quality graphics with an underlying resolution of 4096 x 3120.

Applications Modules

The optional SimGauss module simulates nonlinear differential equations, state-space

and discrete systems. It's powerful, flexible and supports state/parameter vectors for increased speed and ease. Modules are also available for numerical optimization, solving simultaneous nonlinear equations, generating basic descriptive statistics, maximum likelihood, logit, probit, loglinear analysis, and others.

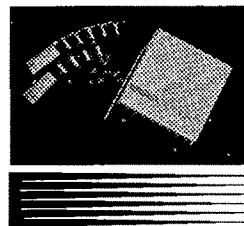
Economical

For just \$495 you get the basic system including a library of mathematical functions and publication quality graphics.

Guaranteed

If you don't agree that GAUSS is the speediest and most versatile of the mathematical programming languages, simply return it for a full refund. No questions asked. Call (206) 631-6679 to order.

GAUSS-faster, easier, smarter... guaranteed.



System includes 700 page manual and 11 program and application disks for IBM PC-XT-PS/2 & compatibles with math coprocessor. 386 protected mode version also available.

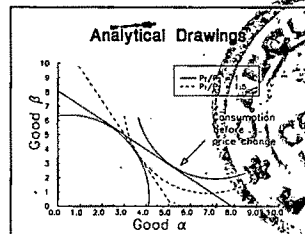
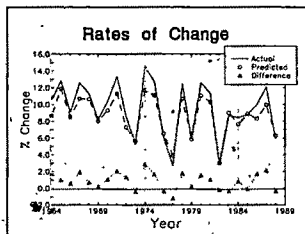
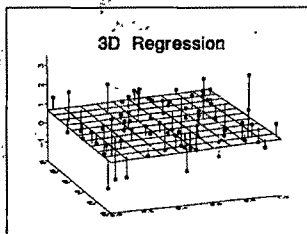
GAUSS™

BY APTECH SYSTEMS, INC.

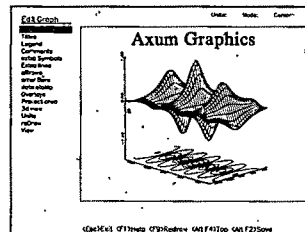
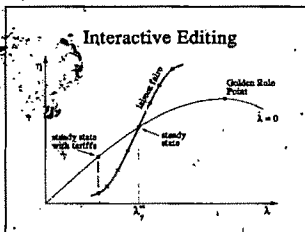
26250 - 196th Place SE, Kent Washington 98042

Phone (206) 631-6679 FAX (206) 630-1220

CALL OR WRITE FOR FREE BROCHURE



Data		Levels		F1		F2		F3	
Cell	1	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Column	2	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Row	3	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Page	4	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Worksheet	5	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
File	6	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Format	7	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Tools	8	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Window	9	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
Help	10	20.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00



Amazing graphics made simple!

Here is what Axum users are saying:

"Fantastic..." "In a matter of minutes I took my Lotus data, created a graph, and used it in WordPerfect..." "Axum is exactly what I've been waiting for." "I created publication-quality graphs without even using the manual..." "...other graphics packages are Stone Age compared with Axum!"

Introducing Axum...

Axum, a new technical graphics and data analysis program for PCs, combines the ease-of-use of the best-selling business graphics packages with the advanced capabilities of expensive mainframe technical graphics programs.

Even if you have never used a graphics program before, with Axum you will be analyzing and graphing your data within minutes.

Publication-quality technical graphs

Take advantage of Axum's extraordinary display capabilities.

- Produce publication-quality, color graphs
- Choose from numerous graph types: 2D, 3D, Contour & more
- Choose from 21 fonts including Greek and scientific; use unlimited super and subscripts
- Graph arbitrary 2D and 3D functions
- Interactively create and customize

- Place multiple graphs on a page: rotate, shrink, or expand
- Automatic regression plots
- Scale plots to multiple y axes and logarithmic axes
- Change the viewpoint of your 3D graph
- Features ultra-high-resolution output from GraphiC™ by Scientific Endeavors

State-of-the-art data editor

Use Axum to perform both simple and advanced data analysis with ease.

- Easily import, edit and transform data
- Explore and analyze your data quickly and easily using multiple, interrelated "data sheets"
- Use unlimited-sized data sets
- Do statistics, curve fitting, and smoothing
- Create macros using Axum's built-in advanced programming language
- Choose from over 100 functions and operators for manipulating and transforming your data

Works with the software you already have

- Use your graphs-in WordPerfect, Word, Ventura, PageMaker
- Export to GEM, Lotus PIC, HPGL, Tektronix, and Encapsulated PostScript files
- Import ASCII, Lotus, and dBase files
- Use virtually any monitor, printer (including PostScript) and plotter

60 day money back guarantee

Try Axum without risk. If you are not completely satisfied, return it within 30 days for a refund.

Axum®

Call toll free for FREE brochure

1-800-548-5653 ext. 030

TRIMETRIX

444 NE Ravenna Blvd Suite 210-AE

Seattle, WA 98115

TEL: WA 206-527-1801

FAX: 206-522-9159

For the IBM PC, XT, AT, PS/2 and compatibles